Thesis for doctoral degree (Ph.D.)
2019

Next generation sequencing to find genetic risk factors in familial cancer

Jessada Thutkawkorapin

Karolinska Institutet

From Molecular Medicine and Surgery
Karolinska Institutet, Stockholm, Sweden

NEXT GENERATION SEQUENCING TO FIND GENETIC RISK FACTORS IN FAMILIAL CANCER

Jessada Thutkawkorapin

Karolinska Institutet

Stockholm 2019

Published by Karolinska Institutet.
Printed by E-Print AB 2019

ISBN 978-91-7831-484-3

Next generation sequencing to find genetic risk factors in familial cancer

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Jessada Thutkawkorapin

Principal Supervisor:
Emma Tham
Karolinska Institutet
Department of Molecular Medicine and Surgery
Clinical Genetics

Co-supervisor(s):
Annika Lindblom
Karolinska Institute
Department of Molecular Medicine and Surgery
Neurogenetics

Daniel Nilsson
Karolinska Institute
Department of Molecular Medicine and Surgery
Rare disease

Opponent:
Esa Pitkänen
European Molecular Biology Laboratory (EMBL),
Heidelberg, Germany
University of Helsinki, Finland

Examination Board:
Bengt Persson
Uppsala University
Department of Cell and Molecular Biology

Tobias Sjöblom
Uppsala University
Department of Immunology

Teresita Diaz de Ståhl
Karolinska Institute
Department of Oncology-Pathology

ABSTRACT

In 2015, Cancer is the second leading cause of death worldwide. Genetic predisposition in familial cancer cases is largely unexplained. At the same time, rapid development in sequencing technology results in an unprecedented increase in the amount of whole exome- and whole genome sequencing data. The studies in this thesis take advantage of the technology and explore possibilities to identify genetic factors behind cancer development.

In **paper I**, we identified 12 novel non-synonymous single nucleotide variants, which were shared among 5 affected members of a family with gastric- and rectal cancer. The mutations were found in 12 different genes; *DZIP1L*, *PCOLCE2*, *IGSF10*, *SUCNR1*, *OR13C8*, *EPB41L4B*, *SEC16A*, *NOTCH1*, *TAS2R7*, *SF3A1*, *GAL3ST1*, and *TRIOBP*. None of the mutations was suggested as a high penetrant mutation We propose this family, suggested to segregate dominant disease, could be an example of complex inheritance.

In **paper II**, we identified a pathogenic variant in *PTEN* in a patient with a Cowden syndrome. We confirmed a pathogenic variant in *PMS2* found in one of the samples suggested by another study. In addition, the study proposed 3 candidate missense variant in known cancer susceptibility genes (*BMPR1A*, *BRIP1* and SRC), 3 truncating variants in possibly novel cancer genes (*CLSPN*, *SEC24B* and *SSH2*), 4 candidate missense variants (*ACACA*, *NR2C2*, *INPP4A* and *DIDO1*), and 5 possible autosomal recessive genes (*ATP10B*, *PKHD1*, *UGGT2*, *MYH13* and *TFF3)*.

The study in **paper III** was to provide a comprehensive local reference database of 1,000 whole genome sequenced Swedish individuals. The samples were selected by principal component analysis from the Swedish Twin Registry (n=942) and The Northern Sweden Population Health Study (n=58). The result illustrated that the genetic diversity within Sweden is substantial compared with the diversity among continental European populations, confirming the importance this database.

The aim of **paper IV** was to identify combinations of both known and unknown cancer processes in humans based on the integration of base substitution-, copy number variation-, structural rearrangement- and microsatellite instability profile in 74 whole genome sequencing tumor-normal pairs from The Cancer Genome Atlas project (TCGA). The results illustrated correlated mutational structure both between and within mutation types, suggesting integrating profiles of several mutation types can enhance accuracy in mutational patterns discovery.

In conclusion, advancement in sequencing- and computational technology demonstrated its capability in identifying cancer causative mutations, proposing candidate genes, providing infrastructure for medical research, as well as visualizing processes underlying cancer development.

LIST OF SCIENTIFIC PAPERS INCLUDED IN THE THESIS

I. **Exome sequencing in one family with gastric- and rectal cancer**.

Thutkawkorapin J, Picelli S, Kontham V, Liu T, Nilsson D, Lindblom A.

BMC genetics 2016, 17(1):41

II. **Exome sequencing in 51 early onset non-familial CRC cases**.

Thutkawkorapin J, Lindblom A, Tham E.

Mol Genet Genomic Med 2019:e605.

III. **SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population**.

Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kahari AK, Lundin P, Che H, Thutkawkorapin J, Eisfeldt J, Lampa S, Dahlberg M, Hagberg J, Jareborg N, Liljedahl U, Jonasson I, Johansson A, Feuk L, Lundeberg J, Syvanen AC, Lundin S, Nilsson D, Nystedt B, Magnusson PK, Gyllensten U.

European journal of human genetics : EJHG 2017, 25(11):1253-1260.

IV. **pyCancerSig: subclassifying human cancer with comprehensive single nucleotide, structural and microsattelite mutation signature deconstruction from whole genome sequencing**

Thutkawkorapin J, Eisfeldt J, Tham E, Nilsson D.

Manuscript (2019)

ADDITIONAL SCIENTIFIC PAPERS

Listed in chronological order

- **Linkage analysis revealed risk loci on 6p21 and 18p11.2-q11.2 in familial colon and rectal cancer, respectively.**

  von Holst S, Jiao X, Liu W, Kontham V, Thutkawkorapin J, Ringdahl J, Bryant P, Lindblom A.

  European journal of human genetics : *EJHG* 2019.

- **Two novel colorectal cancer risk loci in the region on chromosome 9q22.32**.

  Thutkawkorapin J, Mahdessian H, Barber T, Picelli S, von Holst S, Lundin J, Valle L, Kontham V, Liu T, Nilsson D, Jiao X, Lindblom A.

  Oncotarget 2018, 9(13):11170-11179.

- **Cancer risk susceptibility loci in a Swedish population.**

  Liu W, Jiao X, Thutkawkorapin J, Mahdessian H, Lindblom A.

  Oncotarget 2017, 8(66):110300-110310.

- **PHIP - a novel candidate breast cancer susceptibility locus on 6q14.1**

  Jiao X, Aravidis C, Marikkannu R, Rantala J, Picelli S, Adamovic T, Liu T, Maguire P, Kremeyer B, Luo L, von Holst S, Kontham V, Thutkawkorapin J, Margolin S, Du Q, Lundin J, Michailidou K, Bolla MK, Wang Q, Dennis J, Lush M, Ambrosone CB, Andrulis IL, Anton-Culver H, Antonenkova NN, Arndt V, Beckmann MW, Blomqvist C, Blot W, Boeckx B, Bojesen SE, Bonanni B, Brand JS, Brauch H, Brenner H, Broeks A, Bruning T, Burwinkel B, Cai Q, Chang-Claude J, Collaborators N, Couch FJ, Cox A, Cross SS, Deming-Halverson SL, Devilee P, Dos-Santos-Silva I, Dork T, Eriksson M, Fasching PA, Figueroa J, Flesch-Janys D, Flyger H, Gabrielson M, Garcia-Closas M, Giles GG, Gonzalez-Neira A, Guenel P, Guo Q, Gundert M, Haiman CA, Hallberg E, Hamann U, Harrington P, Hooning MJ, Hopper JL, Huang G, Jakubowska A, Jones ME, Kerin MJ, Kosma VM, Kristensen VN, Lambrechts D, Le Marchand L, Lubinski J, Mannermaa A, Martens JWM, Meindl A, Milne RL, Mulligan AM, Neuhausen SL, Nevanlinna H, Peto J, Pylkas K, Radice P, Rhenius V, Sawyer EJ, Schmidt MK, Schmutzler RK, Seynaeve C, Shah M, Simard J, Southey MC, Swerdlow AJ, Truong T, Wendt C, Winqvist R, Zheng W, kConFab AI, Benitez J, Dunning AM, Pharoah PDP, Easton DF, Czene K, Hall P, Lindblom A.

  Oncotarget 2017, 8(61):102769-102782.

CONTENTS

LIST OF ABBREVIATIONS

1000G 1000 Genomes project

CNV Copy number variation

CRC Colorectal cancer

DNA Deoxyribonucleic acid

ExAC Exome Aggregation Consortium

GATK Genome Analysis Toolkit

gnomAD Genome Aggregation Database

HP Hyperplastic polyps

IGV Integrative Genomics Viewer

Kbp Kilobase pairs

MMAF Maximum minor allele frequency

MSI Microsatellite instability

NMF Non-negative matrix factorization

NSPHS Northern Sweden Population Health Study

PCA Principal component analysis.

PCR Polymerase chain reaction

RNA Ribonucleic acid

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

STR Swedish Twin Registry

SV Structural variant

TA Tubular adenoma

TCGA The Cancer Genome Atlas

TVA Tubulovillous adenoma

WGS Whole genome sequencing

1 INTRODUCTION

Deoxyribonucleic acid (DNA) is the hereditary material that contains all the necessary information to build and maintain an organism. This information can be inherited from one generation to the next. One of the basic mechanisms of DNA is that DNA is transcribed into messenger ribonucleic acid (mRNA), and then mRNA is translated in protein to perform its function in the cell. A recent study showed that 25.3% of DNA cannot be transcribed (Djebali et al., 2012), called intergenic regions. The transcribed regions consist of protein-coding genes or non-protein-coding genes (encoding RNA transcripts).

During the transcription process of a protein-coding gene, the gene is transcribed, making a copy of itself in the form of precursor mRNA (pre-mRNA). Pre-mRNA is an immature single strand of mRNA. There are two segments in pre-mRNA, exons and introns. Introns are removed during splicing processes, while exons are retained in the final mRNA. Only 1.1% of human genome is protein-coding exons (Venter et al., 2001).

The nucleotide sequence in the mRNA is read by ribosomes in a sequence of nucleotide triplets, called codons. A three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. The translation starts at the start codon, a triplet of AUG, and keeps translating codons from the nucleic acid sequence until it reaches the stop codon, a triplet of either UAA, UAG, or UGA.

Besides classification from a transcription perspective, the regions can also be classified based on regulatory effects. A promoter is a region located upstream near the transcription start site of a gene on the same strand. Its role is to initiate transcription of the gene. An enhancer is a region of DNA that can be bound by transcription activator to activate transcription of a gene and can be located on the same or on a different strand (Maston et al., 2006, Blackwood and Kadonaga, 1998).

1.1 TYPE OF VARIANTS

Genetic variation can be divided into three categories according to the size and type of the variation: small-scale sequence variation (less than 1Kbp), large-scale structural variation (more than 1Kbp) (Abbs et al., 2004), and numerical variation (whole chromosomes or genomes).

Small-scale sequence variation can be divided into two sub-categories, single base-pair substitutions, and insertions or deletions (indel). The variants can be caused by translesion synthesis (Waters et al., 2009), defect DNA repair (Lieber, 2010), and mutagens (Papavramidou et al., 2010). A single base-pair substitution is a change in DNA sequence in which one base pair is altered. If the variant occurs in an exonic region, it can have direct effect on the coding protein downstream in many ways: missense, stop-gained, stop-lost, inframe indel or frameshift indel. A silent variant, or a synonymous variant, is a variant that doesn't change the protein product but the variant can be pathogenic if the variant becomes a splicing motif promoting exon skipping or removes a splice site. A missense variant is a variant that

change the protein product but the length is preserved. A stop-gained variant, or a nonsense variant, is a change that results in a premature stop codon, leading to a shortened protein, and possibly resulting in nonsense-mediated mRNA decay. A stop-lost variant results in a change in at least one base of the stop codons, resulting in an abnormal elongated protein. An indel is a change in the nucleotides sequence, which can be an insertion- or a deletion of nucleotides. This results in a net change in total number of nucleotides. An inframe variant is a change in which triplets are gained or missing. It does not cause a disruption of the translational reading frame. A frameshift variant is a change in which the number of inserted- or deleted nucleotides is not a multiple of three. This causes a disruption of the translation reading frame.

At a larger scale, variants can be divided into unbalanced and balanced events. Unbalanced events happen when the change in DNA content results in extra copies or missing DNA material. These events include structural duplication and structural deletion resulting in increasing or decreasing amounts of genetic material. This may increase or decrease activities at RNA and/or protein levels. On the other hand, balanced events result in the same amount of genetic material. These events include structural inversion and translocation. Fusion transcripts from such events may cause cancer e.g. lymphoma (Li et al., 1999, Streubel et al., 2003) and thyroid cancer (Klemke et al., 2011)

1.2 MENDELIAN PEDIGREE PATTERNS

Usually, the expression of any human phenotype depends on many genes and environmental factors. But it is also possible for a phenotype to be expressed with only a particular genotype at one locus, given the normal genetic and environmental background. These phenotypes are called mendelian. Common patterns include mono-allelic-, bi-allelic, and de novo disorder.

Mono-allelic disorder is the simplest pattern, especially if it's high penetrance rare disorder, usually with one of the parents carrying the disease allele and have the disease. There is a 50% probability for each of the affected sibling to have the disorder.

Bi-allelic and de novo inheritance patterns are hard to be differentiated from the pedigree, for the high penetrance disorder, as none of the parents is affected. However, genetically, they are different. In bi-allelic, both of the parents carry the disease allele, while, in de novo, none of them has the allele. Mathematically, in bi-allelic cases, siblings of the affected has 25% probability to have the disease. In de novo, the risk is varied depending on when the pathogenic event was triggered.

In cancer, if one of the parents carrying a pathogenic variant in a high-penetrance tumor suppressor gene, there is 50% risk for each of the children to inherit this variant. The child with the disease allele will have every cell in their body with this variant. If there is a seond-hit event, another pathogenic event triggered in the other allele of the same gene, the tumor suppressor mechanic will lose its function, thus cancer start to develop.

1.3 CHARACTERISTICS OF MENDELIAN INHERITANCE

There are various complications that often disguise a basic Mendelian pattern.

The penetrance of a phenotype is defined as the probability that a person who has the genotype will express the phenotype. From definition above, a dominant phenotype is expressed in a heterozygous individual, and should show 100% penetrance. In reality, 100% penetrance is the more unusual phenomenon.

Late-onset diseases are particular important cases of reduced penetrance. The diseases are age-related in the sense that the phenotype is not expressed until adult-life. The delayed effect might be caused by slow accumulation of somatic mutations. Good examples of inherited diseases with delayed onset are the familial cancer syndromes, where the affected individual inherits the variant from one of the parents and has the second hit later in life (Cavenee et al., 1983).

Common recessive conditions can give a pseudo-dominant pedigree pattern. If a phenotype is common in the population, there is a high probability that it may be brought into the pedigree by two or more individuals independently. Consanguinity can cause the same phenomenon.

There are also situations when individuals carrying the same genotype express a non-binary phenotype or different phenotypes (Konno and Silm, 2001). These situations are called variable expressivity. Other genes, environmental factors or pure chance may contribute to the variability of phenotypes.

Certain human phenotypes are autosomal dominant but they are expressed only when the genotype is inherited from a parent of one particular sex. The genes that contribute to such effects are called imprinted genes.

Male lethality may complicate X-linked pedigrees as the affected die before birth. Thus, the variants can only be passed to half of their daughters but none to their sons.

De novo variants often complicate pedigree interpretation, and can be mosaic. A de novo variant is a variant that is present for the first time in a family. None of the parents are affected or carriers. An example of this is when a healthy couple with no relevant family history have a child with severe abnormalities. The mode of inheritance might be autosomal recessive, de novo autosomal dominant, X-linked recessive (if the child is male), or purely environmental factors. This makes it hard for the interpretation and for estimating the recurrence risk.

Phenocopy is a phenotype that mimics the disease phenotype but is caused by other factors (Goldschmidt, 1949). If phenocopies cannot be identified before designing the study, they can lead to wrong hypotheses and, eventually, incorrect findings.

1.4 GENETIC DISEASES

Abnormalities in human genetics can manifest itself regardless of age, sex, family background. It can affect growth and childhood development (Byard, 1994, Bobadilla et al., 2002, Malt et

al., 2013). It can also be delayed and have the effect in adults, as in cancer (Cavenee et al., 1983).

1.4.1 Inherited colorectal cancer

Colorectal cancer (CRC) is the third most common cancer type worldwide. The estimated risk for those, who have first-degree relatives diagnosed with CRC, is increased by two to four fold (Johns and Houlston, 2001). Around 7% of CRC cases are diagnosed at an age less than 50, while 20% of the cases have at least one first-degree relatives with CRC (Burt, 2000). However, less than 5% of familial cases are identified as known cancer syndromes (Syngal et al., 2005).

1.4.2 Known CRC syndromes

Lynch syndrome, which can also be called hereditary nonpolyposis colorectal cancer (HPNCC), is an inherited autosomal dominant cancer syndrome, that contributes to an increased risk of several types of cancers, including colorectal-, endometrial-, ovarian-, gastric-, upper urinary tract-, and biliary tract cancer (Kohlmann and Gruber, 1993). The disease is caused by pathogenic variants in DNA mismatch repair (MMR) genes. There are four genes known to cause Lynch syndrome; *MLH1*, *MSH2*, *MSH6*, and *PMS2*, with life-time risk 46%, 35%, 20%, and 10% respectively (Moller et al., 2015). Lynch syndrome is accounted for 1-3% of all colorectal cancer cases (Burt, 2007)

Familial adenomatous polyposis (FAP) is an autosomal dominant disease, caused by a pathogenic variant in the adenomatous polyposis coli (*APC*) gene. In FAP, hundreds to thousands of adenomatous polyps form in the rectum and colon. The polyps are initially benign but they will be transformed into cancer if they are not identified and treated at an early stage (Half et al., 2009). FAP accounts for less 1% of CRC cases (Reed and Neel, 1955, Alm, 1975)

MUTYH-associated polyposis (MAP) is an autosomal recessive form of inherited polyposis. It is caused by biallelic pathogenic variants in *MUTYH* (Nielsen et al., 2012). The number of polyps are between ten to a few hundred (Nielsen et al., 2011, Grover et al., 2012). If the polyps are not identified or left untreated, the lifetime risk of developing CRC is between 43% to 100% (Sampson et al., 2003, Sieber et al., 2003, Gismondi et al., 2004, Farrington et al., 2005, Lubbe et al., 2009).

1.4.3 Pathways to colorectal cancer

CRC can be caused by environmental factors, genetic changes or epigenetic alterations. Examples of environmental factors are obesity (Le Marchand et al., 1997, Slattery, 2004), and food (Agnoli et al., 2013). Genetic and epigenetic alteration can initiate the transformation of normal colon tissue into adenoma, and finally into cancer (Fearon and Vogelstein, 1990).

1.4.3.1 Chromosomal instability pathway

Most sporadic CRC cases fall into the chromosomal instability pathway category due to several loss of heterozygosis (Lin et al., 2003) and chromosomal aberrations (Leary et al., 2008). Most

of these tumours have somatic mutations in *APC*. *APC* not only controls how often a cell divides but also controls the number of chromosomes during cell division (Fodde et al., 2001, Powell et al., 1993).

1.4.3.2 Microsatellite instability pathway

One form of genomic instability is hypermutation caused by often caused by the inactivation of DNA mismatch repair (MMR) systems. The function of the MMR system is to identify mismatches in the DNA and to direct the repair machinery (Boland et al., 1998). The dysfunction of MMR system results in errors during DNA replication, which can be measured by analysis of different sizes of microsatellite alleles (Peltomaki et al., 2001), so-called microsatellite instability. Most CRC with MSI is caused by somatic methylation of the *MLH1* promoter and is associated with a CpG Island Methylator Phenotype (CIMP) (Cancer Genome Atlas, 2012). A well-known cancer syndrome caused by germline pathogenic variants in MMR genes is Lynch syndrome.

1.4.3.3 Epigenetics alterations pathway

Pathological epigenetic changes are emerging factors disrupting gene function (Egger et al., 2004). The changes include histone modification, DNA hypo- and hyper methylation, and loss of imprinting. The changes lead to dysfunctions of cell cycle regulation, apoptosis, angiogenesis, DNA repair, invasion and adhesion.

1.4.3.4 Other pathways

MicroRNAs (miRNAs) play an important role in RNA silencing and post-transcriptional regulation of gene expression (Ambros, 2004, Bartel, 2004). Recent studies found that altered expression of 13 miRNAs may be associated with regulatory action in RAS pathway (Bandres et al., 2006) in CRC patients.

1.5 METHODS

1.5.1 Genetic linkage analysis

Genetic linkage analysis is a powerful technique traditionally used in monogenic diseases to identify high-risk predisposing genes such as *APC* (Bodmer et al., 1987), *MLH1* (Lindblom et al., 1993), and *MSH2* (Peltomaki et al., 1993). It is based on the observation that alleles residing physically close on a chromosome tend to be inherited together during meiosis. This type of analysis requires a few large families with many small families believed to have the same phenotype suggesting the same causative gene. The result from the analysis is the logarithm of the odds (LOD) score. A LOD score of 3 or more is generally accepted as an indication that 2 loci are linked.

1.5.2 Association study

Low-risk variants cannot be identified using linkage analysis since they rarely result in pedigrees with many affected (Risch and Merikangas, 1996). However, it can be done using

association studies using numerous samples. Recently, several new susceptibility loci have been discovered by various association studies, often including many thousands of cases and controls (Peters et al., 2015, Zhang et al., 2014, Wong et al., 2013).

1.6 HIGH THROUGHPUT SEQUENCING ANALYSIS (MASSIVELY PARALLEL SEQUENCING)

Cheaper and cheaper cost per reaction of DNA sequencing introduced by massively parallel sequencing (MPS) (Mardis, 2008) allows molecular research to be performed at base-pair resolution. The MPS applications include genome sequencing and resequencing, transcription profiling (RNA-Seq), DNA-protein interactions (ChIP-Seq), and epigenome sequencing (de Magalhaes et al., 2010).

Resequencing is DNA sequencing in an organism for which a reference genome is available and used. In human genome resequencing studies, whole genome- or targeted sequencing can be performed. Whole genome sequencing (WGS) denotes the sequencing of the entire genome, while targeted-, sometimes called capture-based, only focuses on specific regions, such as coding regions, gene panels, or custom regions (Grody et al., 2013).

The obvious advantage of WGS over the targeted approach is the amount of data: entire genome is sequenced compared to only protein coding regions, which is around 1% of the genome. Moreover, WGS approach gives higher SNP detection sensitivity (Meynert et al., 2014) (Fang et al., 2014). On the other hand, the targeted approach has economic advantages, not only for the sequencing, but also regarding the storage, and computational resources, thus ability to sequence more deeply for low fraction mosaic variants in e.g. tumor material.

The targeted approach that focuses on coding regions, usually called whole exome sequencing, has been used to identify and confirm various novel disease candidate genes in CRC studies, such as *EIF2AK4* (Zhang et al., 2015), *MLL3* (Li et al., 2013), *NTHL1* (Weren et al., 2015), *FAN1* (Segui et al., 2015), *CDKN1B*, *XRCC4*, *EPHX1*, *NFKBIZ*, *SMARCA4*, *BARD1* (Esteban-Jurado et al., 2015), and in *POLD1* and *POLE* genes (Chubb et al., 2015, Valle et al., 2014).

A typical workflow consists of library preparation, then sequencing by the instrument, followed by quality control, alignment & variant detection, and then data analysis.

1.6.1 Library preparation

In general, the library preparation steps involve shearing the DNA sequence into small fragments, with insert size varying between 100 base-pairs to several kilo base-pairs, depending on amount of input DNA, technology and the preparation protocol. Then, the fragments are ligated with adaptors at 5' and 3'. Now, the ligated fragments are ready for cluster generation and sequencing.

1.6.2 Quality control

Ideally, researchers would expect to have sequencing data with exactly the same content as the human DNA, having reads mapping evenly and having 50/50 paternal/maternal alleles. Unfortunately, there are several factors influencing the quality of the data, for example, contamination, DNA quality, limitation caused by the technology, and technical errors. In order to evaluate the reliability of the data, quality control of sequencing reads has to be performed.

1.6.2.1 Average depth

Average depth, sometimes called depth of coverage, is defined by summing depth of all target sequencing bases and then dividing by the number of bases. At one base position, higher depth means higher statistical significance of base calling.

1.6.2.2 Percentage of mapping with at least X depth

This assessment usually comes together with calculating average depth. As the name "average depth" imply, not all the target DNA is covered evenly. One reason is that there are low complexity or non-unique regions in the DNA that is difficult to map or unmappable (Figure 1). Percentage of mapping with at least X depth, when X represent number of depths with good enough statistical significance, can report of the sequencing data with reliable base calling quality.

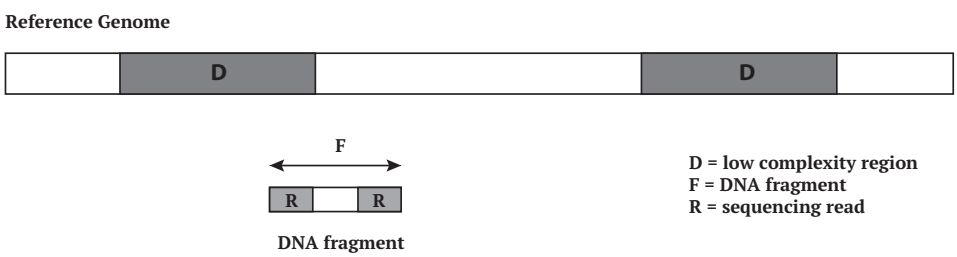

Figure 1. Illustration of unmappability. If there are two regions in the DNA (D) that are identical and size of F is far smaller than D, during read mapping steps, reads that cannot be uniquely mapped to the reference genome is not guarantee to be mapped evenly. In the worst case, it's possible 100% of these reads will be mapped to only one D leaving the other D with no mapping at all. Thus, one D with two times average and zero depth in the other.

1.6.2.3 GC content

Base calling error has been shown to not be equally distributed with all base substitutions. Moreover, the error rate become higher toward the end of the reads (Dohm et al., 2008). Statistically, the errors are frequently preceded by base G. The most common error is A > C substitution and the least is C > G substitution. Unusual distribution of GC content can suggest sequencing bias during library preparation and base calling.

1.6.2.4 Duplicate reads

Traditionally, PCR amplification is required as one part of the library preparation. However, this step can introduce PCR duplicates, sequencing reads with exactly the same DNA fragment. The PCR duplicates can result in false detection of copy number variation. Moreover, if there are errors in the reads, they may propagate and result artefact during variant calling process.

1.6.3 Alignment and variant detection.

The goal of this step is to convert sequencing reads into a list of variants. In general, this step involves read alignment and variant calling. Additional steps can be included to improve the quality of variant detection, such as MarkDuplicate, Indel recalibration, Variant recalibration, and variant quality score recalibration (DePristo et al., 2011, Van der Auwera et al., 2013). Types of variants that can be detected include base substitution, small insertion, small deletion, copy number variation, and structural rearrangement.

1.6.3.1 Read alignment

Read alignment is a group of processes to correctly align reads back to the human reference genome. The Genome Reference Consortium (GRC) the human reference genome, and several organizations provide interfaces and additional resources, e.g. the University of Santa Cruz (UCSC) (Raney et al., 2011, Rosenbloom et al., 2012). The alignment processes involved encompass read mapping, indel realignment, and base recalibration. Read mapping is to map raw read data to the reference. There are many bioinformatics software available to do this task, including Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009, Li and Durbin, 2010), YOABS (Galinsky, 2012), CUSHAW2 (Liu and Schmidt, 2012), SOAP (Li et al., 2009b, Luo et al., 2013), and Stampy (Lunter and Goodson, 2011). They have different strengths based on different criteria, which are processing time, memory usage, read size, sequencing instrument, license type, and multi-threading. The next step after mapping reads is to realign indels. Identification of indels based on independently mapped reads may lead to incorrect indels and SNPs, especially if indels are at the end of the reads. The indel realign process is to use previously mapped reads altogether to determine indels. Among available tools, Genome Analysis Toolkit (GATK) is one of the most widely accepted (Van der Auwera et al., 2013, DePristo et al., 2011, McKenna et al., 2010). The final step of reads alignment is base recalibration. Quality scores of individual bases in the reads heavily influence the algorithm in variants calling and the estimated scores provided by the sequencing machines are subject to various sources of systematic technical errors. The role of base recalibration is to adjust the quality scores based on the data and known variants.

1.6.3.2 Single nucleotide variant and small insertion-deletion calling

The greatest challenge in this step is to minimize the number of false positives and false negatives. In general, the information that germline variant callers use to identify the zygosities are base quality scores and base ratios. Several software programs have been developed, such as GATK (McKenna et al., 2010), BCFtools (Li et al., 2009a), FreeBayes (Erik Garrison,

2012), MuTect2 (Cibulskis et al., 2013), VarScan2 (Koboldt et al., 2012), ExScalibur (Bao et al., 2015), Fermikit (Li, 2015), BAYSIC (Cantarel et al., 2014), FAVR (Pope et al., 2013), and VarDict (Lai et al., 2016). Their differences are size of called indels, computational time, memory usage, multi-threading, accuracy, platform specific, and read-depth.

1.6.3.3 Copy number variant calling

There are a few computational ways to detect copy number events. One way is to use paired reads or split reads as evidence. For example, if a paired read is mapped to a coordinate far apart from the position expected given insertion size, it can be used as an evidence suggesting structural deletion. The tools with this method include TIDDIT (Eisfeldt et al., 2017) and MANTA (Chen et al., 2015). Another is to use sequencing coverage across the genome. This assumes that the depth is roughly equal across the genome. Any regions with average depth significant lower or higher than expected suggest structural deletion and structural duplication events respectively. The tool employing this method include CNVnator (Abyzov et al., 2011). For exome sequencing data, where the depth is uneven, CNV detection can be done using another sample, or a collection of samples, as a reference. The tools that use this method include ExomeDepth (Plagnol et al., 2012) and XHMM (Fromer et al., 2012).

1.6.3.4 Structural rearrangement

The method detecting structural rearrangement include paired reads as an evidence. Thus, TIDDIT (Eisfeldt et al., 2017) and MANTA (Chen et al., 2015), employing this method in CNV detection, can also detect structural rearrangement.

1.6.4 Data analysis processes

To make list of variants become actionable knowledge, variant need to be annotated to provide their biological significance, which helps in filtering and prioritizing disease-causing variants. The information that can be annotated to the variants includes, frequency-, structural-, prediction- and evidence-based data. Besides annotation, visualization plays an important role in the interpretation of the result as it turns the information in computer readable format into a human readable visual representation.

Typically, an exome sequencing in one sample can identify up to 30,000 variants (Kassahn et al., 2014). To narrow them down to a list of a few candidate variants, variant annotation is performed to integrate evidences from different sources to predict variant significance. There are commercial software solutions that are packaged with sequencing instruments, such as VariantStudio (Illumina), IonReporter (Life Technologies), Geneticist Assistant (Softgenetics), and Expressionist (GeneData). Open-source software, such as ANNOVAR (Wang et al., 2010), the Ensembl Variant Effect Predictor (McLaren et al., 2016), and snpEff (Cingolani et al., 2012), can add basic information, gene name, transcripts, and regulatory regions, to the variants. PFAM (Xu and Dunbrack, 2012) and SMART (Letunic et al., 2012) can predict functional significance with respect to known protein domains. It is important to keep in mind that silent variants, which don't change the translated protein downstream, sometimes can alter

splicing (Arnold et al., 2009). Non-coding variants, regulatory regions, and splice sites, can be annotated by SPIDEX (Jagadeesh et al., 2019), ENCODE (Fratkin et al., 2012), and FANTOM (Kawaji et al., 2009, Kawaji et al., 2011). *In silico* pathogenicity prediction is used to predict the probability of a sequence alteration to affect protein function. The predictors were developed based on three strategies, evolutionary conservation, structural or biophysical properties, and machine learning techniques. Some popular predictors are SIFT (Kumar et al., 2009, Sim et al., 2012), PolyPhen2 (Adzhubei et al., 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2010), Phylop (Cooper et al., 2005), GERP++ ++ (Davydov et al., 2010), and CADD (Kircher et al., 2014).

There are several databases that host previous findings and clinical evidence. The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001) contains a range of molecular variation: SNPs, indels, microsatellites, multinucleotide polymorphisms (MNPs), heterozygous sequences, and named variants. DGV and dbVar (Lappalainen et al., 2013) contain genomic structural variations. The database of Genotypes and Phenotypes (dbGaP) (Mailman et al., 2007, Tryka et al., 2014) archives and distributes the data and results from studies that have investigated the interaction of genotype and phenotype in Humans. OMIM (Hamosh et al., 2005) has knowledge-based information of human gene and genetic disorders. The NHLBI Exome Sequencing Project (ESP) contains exome sequencing data and related phenotype data from populations of heart-, lung- and blood disorders. HGMD (Stenson et al., 2014) is a database of known gene lesions responsible for human inherited disease. Leiden Open Variation Database (LOVD) (Aartsma-Rus et al., 2006) is a tool for Gene-centered collection and display of DNA variations. ClinVar aggregates information about genomic variation and its relationship to human health (Landrum et al., 2018). InSiGHT houses and curates the most comprehensive database of DNA variants re-sequenced in the genes that contribute to gastrointestinal cancer (Thompson et al., 2014). ENIGMA is an international consortium of investigators focused on curating sequence variants in BRCA1, BRCA2 and other known or suspected breast cancer genes. The curated genetics information from InSiGHT and ENIGMA have been routinely incorporated into ClinVar. COSMIC (Forbes et al., 2017) stores and displays somatic variant information. The Cancer Genome Atlas (TCGA) project has applied high throughput technologies to sequence human tumors, and human healthy tissues at the DNA, RNA, protein and epigenetic levels. The Genome Aggregation Database (gnomAD) aggregates and harmonizes exome sequencing data from a wide variety of large-scale sequencing projects (Lek et al., 2016).

Visualization is a very crucial step in high-throughput data analysis as it turns information from computer format that require computer skill to understand into friendly visual presentation. Comma-separated values (CSV) and tab-separated values (TSV) can be exported from most of the annotator software and, then, can be imported to any spreadsheet viewer. The Ensembl Genome Browser (Spudich and Fernandez-Suarez, 2010) and the University of Santa Cruz (UCSC) Human Genome Browser (Kent et al., 2002, Speir et al., 2016) are web-based browsers that integrate various sources of annotations. Integrative genomics Viewer (IGV)

(Robinson et al., 2011) is a stand-alone browser that support array-based and massively parallel sequencing data, and genomic annotations

1.6.5 File formats

Communication between each process in high throughput sequencing analysis has been done in file formats accepted by scientific community.

1.6.5.1 FASTA

A file in FASTA format is a texted-based file with the purpose to store nucleotide sequences or protein sequences. Common usage of FASTA files in Human DNA analysis is for storing human reference sequencing, for example, GRCh37 and GRCh38. The reference FASTA files are used by almost all processes in DNA sequencing analysis.

1.6.5.2 FASTQ

A file in FASTQ format a texted-based file with the purpose to store nucleotide sequences together with its corresponding quality scores. Each sequence represents a sequencing read. Normally, in a FASTQ file, there are 4 lines for one read. The first line is the read's ID. The second line is the nucleotide sequence. The third line starts with "+" and is followed by optional sequence information. And the fourth line is the quality scores. In paired-end sequencing, there usually is a pair of FASTQ files for one sample with equal amount of reads in both files. Sequences with the same ID represent a pair of reads. Generally, files in FASTQ format are considered as complete products from sequencers. The files are then used by an alignment tool to align with a reference genome.

1.6.5.3 SAM

A file in SAM format is texted-based file for storing nucleotide sequences, from the corresponding FASTA files, aligned to a reference genome. The files are the output of an alignment tool. A file in SAM format have a header and an alignment section. The header contains overall information of the file or the sample, such as the genome reference used during the alignment process. The alignment section contains the alignment details. Each line has 11 mandatory tab-limited fields. One line represents one alignment.

1.6.5.4 BAM

A file in BAM format is the binary/compressed version of a SAM file. After an alignment process, a SAM file is then compressed into a BAM file. Files in BAM format are the most common intermediate files used in sequencing analysis. Alignments in the file can be visualized using IGV (Robinson et al., 2011). All of the recalibration processes to improve alignment quality take input in BAM format and also output in BAM format (DePristo et al., 2011). Most of SV callers need input in BAM format (Chen et al., 2015, Abyzov et al., 2011, Eisfeldt et al., 2017). All SNV callers need input in BAM format (DePristo et al., 2011, Li et al., 2009a).

1.6.5.5 VCF

A file in VCF format is a text-based file for storing genetic variations. The file is an output of all SNV callers and some SV callers (Chen et al., 2015, Eisfeldt et al., 2017). The file consists of a header section and a variant section. The header section contains overall information of the file, such as the samples name, format of the genotyping data, and format of the annotation data. The last line of the header section is the columns name of the variant section. Each line in the variant section is in tab-limited format. One line represents one variant. Number of columns in the variant section are varied depending on number of samples in the file. Many annotation tools (Wang et al., 2010, Cingolani et al., 2012, McLaren et al., 2016) have an option to output in VCF format.

1.6.5.6 TSV

A TSV file is a tab-limited text-based file with "tab" as an intent for tabular structure, which later can be imported in to Excel. A few variant callers can output in TSV format (Wang et al., 2010, Abyzov et al., 2011). Even though the TSV file format has its strength in its multi-purpose and its readability compared to VCF, they are not commonly used for storing genetic variations mainly because of their size and their inefficiency in storing complex information, such as multiallelic variants or variant with multiple transcripts.

1.6.5.7 CSV

A CSV file is a comma-limited text-based file with the same intent as TSV to be used as tabular structure. However, it's not popular due to its non-standardized format. Errors can be introduced if there are commas or new lines in its content. The file is sometimes used as an exported intermediate to be imported into Excel.

1.7 CANCER SIGNATURES AND DNA DAMAGE PATTERNS

DNA damage accumulated in our cells results from a combination of exposure to damaging processes and losing the related repair mechanisms. Source of the damage can be either exogenous, such as radiation or chemicals, or endogenous, such as replication errors or routine processes in the cell. Our cells have evolved to have repair mechanisms to handle each type of DNA damage.

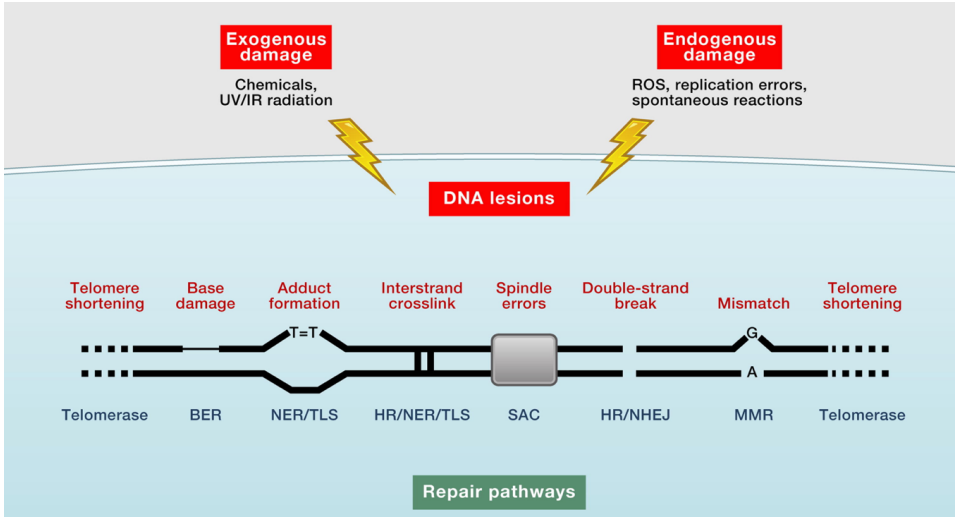

Figure 2. DNA damage and corresponding mechanism. BER, base excision repair; HR, homologous recombination; NER, nucleotide excision repair; NHEJ, nonhomologous end-joining; MMR, mismatch repair; ROS, reactive oxygen species; TLS, translesion synthesis; SAC, spindle assembly checkpoint. [Permission obtained from Elsevier to reuse parts of figure 2 from López-Otín et al. (Lopez-Otin et al., 2013)]

Conventional cancer studies have their focus mainly on identifying cancer-related genes by targeting driver mutation in tumors. Identifying driver mutations can be comparable to "finding a needle in a haystack" as the majority of mutations in tumors are passenger mutations. However, this "haystack" has been shown to be a rich source of information revealing cancer-related mechanisms (Rubin and Green, 2009). Thus, patterns of passenger mutations can be potentially used as a proxy of the mutational processes.

There are studies that have revealed the footprint of cancer mutational processes. For example, in UV light-associated skin cancer, there are patterns of CC:GG > TT:AA double nucleotide substitutions (Pfeifer et al., 2005). In smoking-associated lung cancer, there are patterns of C:G > A:T transversions (Hainaut and Pfeifer, 2001). Biallelic mismatch repair deficiency in *POLE*-mutated cells results in ultra-hypermutated tumors with TCT > TAT and TTT > TGT mutations (Shlien et al., 2015).

1.7.1 Pattern analysis methods

With an exploding amount of whole exome and whole genome sequencing data, together with advances in machine learning technology, a computational approach to characterize tumors based on their base-substitution profile was initiated, capturing mutation signatures (Alexandrov et al., 2013).

Machine learning is a computational technique that lets a computer mimic human behaviour. Known machine learning applications that have been integrated in human daily life include

voice recognition, finger-print recognition, hand-writing recognition, face recognition, and language translation. The strength of machine learning is in its almost-human ability to recognize patterns. It can associate large numbers of observed variables with a set of labels or outcomes that are not necessary exactly the same. With this strength, a machine learning method can easily take advantage of the large amount of passenger mutations and identify associated mutational processes.

1.7.1.1 Supervised learning

Supervised learning is a branch of machine learning that needs pairs of observed variables and outcomes (or sample labelling). The goal of supervised learning is to develop a predictor model. Real life applications include face recognition and finger-print recognition. A known genetics application that used this approach is HRDetect (Davies et al., 2017), which can identify *BRCA1*/*BRCA2*-deficient tumors with 98.7% sensitivity.

1.7.1.2 Unsupervised learning

This is another branch of machine learning that only needs set observed variables (without labels). One of the goals is to cluster samples with a similar pattern. This approach is especially useful when we do not know the number of expected clusters. By being clustered, the additional profit of unsupervised learning is in its visualization that the model reduces the number of observed variables down to a level that can be visualized. Real life applications include any application that can suggest similarity, for example, Facebook that can suggest photo tagging or friend adding. Unsupervised learning can identify tumors with similar patterns regardless of if the patterns are already known (Alexandrov et al., 2013) or unknown.

2 MATERIALS AND METHODS

2.1 COHORTS

The cohorts in *paper I* and *paper II* were Swedish colorectal and breast cancer patients recruited through 14 different hospitals from central Sweden. In *paper I*, the cases from family 242 and the cohort of 98 familial colorectal cancer patients were collected through the Karolinska Hospital, Stockholm, Sweden. In *paper II*, the 51 colorectal cancer patients with age of onset less than 40 without family history were recruited through the Department of Clinical Genetics, Karolinska University Hospital Solna or recruited in a nation-wide study, the Swedish Low-risk Colorectal Cancer Study. The familial breast cancer patients, used as a comparison group, in *paper I* and *II* were recruited through the Department of Clinical Genetics, Karolinska University Hospital Solna.

For family 242, the family segregates early onset rectal- and gastric cancer over three generations suggesting a dominant inherited predisposition. In total there were six cases with early-onset rectal cancer and in total at least four cases with gastric cancer. Some family members were affected also with other cancer types; two men had prostate cancer and both died from their disease, one woman had head-neck cancer and died from the disease and another woman had lymphoma and died because of that. Many family members had presented with tubular adenomas and hyperplastic polyps under surveillance. In particular, four family members had lesions, which could be used for coding of affected status in our study. One woman (Co-652) had three large tubulovillous adenomas (TVA), her sister (Co-692) had four tubular adenomas (TA) and 8 hyperplastic polyps (HP), and another sister (Co-657) had 5 large HP. One man (Co-771), whose mother (Co-666) had died from rectal cancer, had rectal cancer. They were all coded as affected in the first linkage analysis, another study before the study in family 242 (Picelli et al., 2008). One woman with gastric cancer (Co-441) and two relatives with rectal cancer (Co-666 and Co-771) were used for the initial exome sequencing study (Figure 3) (Thutkawkorapin et al., 2016).

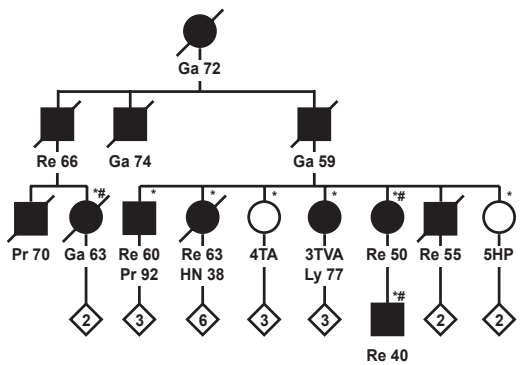

Figure 3. Pedigree of family 242. Ga: Gastric cancer, Re: Rectal cancer, Pr: Prostate cancer, HN: Head and neck cancer, Ly: Lymphoma, TA: Tubular adenomas, TVA: tubulovillous adenomas, HP: Hyperplastic polyps.

In *paper III*, the cohorts were recruited through the Swedish Twin Registry and the Northern Sweden Population Health Study.

In *paper IV*, the cohorts of 8 colorectal- and 66 breast cancer patients were recruited through various studies as parts of The Cancer Genome Atlas project.

2.2 MASSIVELY PARALLEL SEQUENCING

In *paper I, II* and *III*, the sequencing samples were only germline DNA from the patients. In *paper IV*, the data were from both germline and tumor DNA from the patients.

The first 3 patients, one with gastric cancer and two with rectal cancer, from family 242 have been sequenced together with 30 patients from the family breast cancer cohort. They were whole exome sequenced using the SureSelect XT Human All Exon 50 Mb kit on Illumina HiSeq 2000.

The 98 CRC patients from *paper I*, the 51 early-onset CRC patients from *paper II*, and the rest of familial breast cancer patients were whole exome sequenced using the TruSeq PE Cluster Kit v3 on Illumina HiSeq 2000.

In *paper III*, the 1000 samples were whole genomes sequenced using TruSeq DNA PCR free sample preparation kit on Illumina Hiseq X.

In *paper IV*, the 74 cases were whole genome sequenced on Illumina platform and whole exome sequenced on Roche and Applied Biosystems platform.

2.3 DATA PREPROCESSING METHODS

All of the data used in these studies were whole genome and whole exome sequencing data. Thus, some of the data were pre-processed in a similar way. The following were methods used in chronological order.

2.3.1 Alignment and variant calling

In *paper I, II* and *III*, the alignment and variant calling processes were performed according to GATK best practice (Van der Auwera et al., 2013). The processes started with aligning raw reads to GRCh37 version of human reference using BWA-MEM (Li and Durbin, 2009). The aligned reads were sorted and indexed using samtools (Li et al., 2009a). Then, duplicated reads were marked using Picard (broadinstitute.github.io/picard). The indels were realigned using GATK RealignerTargetCreator and IndelRealigner. Base quality scores were recalibrated using GATK BaseRecalibrator. These processes produced a bam file for each sample. GATK HaplotypeCaller were used for creating gVCF files. And the joint VCF files were produced using GATK CombineGVCFs and GATK GenotypeGVCF.

In *paper IV*, the alignment and variant calling were processed as a part of TCGA project. The BAM files used in the study were whole genome sequencing data, aligned to GRCh37. The VCF files were called from whole exome sequencing data, aligned to GRCh38, using MuTect2 (Cibulskis et al., 2013). Both bam and VCF files were downloaded from https://portal.gdc.cancer.gov, with a permission obtained from dbGaP

2.3.2 Variant annotation

In *paper I* and *II*, the merged VCF files were annotated using ANNOVAR (cite). The annotation databases included RefSeq gene annotation (O'Leary et al., 2016), dbSNP (Sherry et al., 2001), ClinVar (Landrum et al., 2018), ExAC conservative constraint (Lek et al., 2016). Background allele frequencies are from SweGen (Ameur et al., 2017), ExAC (Lek et al., 2016), gnomAD (Lek et al., 2016), and 1000 Genomes Project allele frequencies (1000 Genomes Project Consortium, 2012), 200Danes (Y. Li et al., 2010), and 249Swedes (http://neotek.scilifelab.se/hbvdb/)). In silico predictors used for predicting pathogenic effects include SIFT (Kumar, Henikoff, & Ng, 2009), PolyPhen2 (Adzhubei et al., 2010), Phylop (Cooper et al., 2005), LRT (Chun & Fay, 2009), Mutation Taster (Schwarz, Rodelsperger, Schuelke, & Seelow, 2010), Mutation Assessor (Reva, Antipin, & Sander, 2011), FATHMM (Shihab et al., 2015), GERP++ (Davydov et al., 2010), and CADD (Kircher et al., 2014).

2.3.3 Maximum minor allele frequency (MMAF)

In *paper II*, maximum allele frequencies from 21 population (SweGen, ExAC, gnomAD, 1000Genomes, 200Danes and 249 Swedes) were used for filtering.

2.3.4 Sanger sequencing

The PCR primers used in *paper I* and *II* were designed using Primer3web (Untergasser et al., 2012) and SimGene Primer3 (Rozen and Skaletsky, 2000). The sequences were visualized and analyzed using FinchTV (http://www.geospiza.com/Products/finchtv.shtml) and CodonCode Aligner (http://www.codoncode.com/aligner/index.htm).

2.3.5 Structural variant calling

In *paper IV*, the SV calling for both tumor and germline WGS data were done using FindSV (https://github.com/J35P312/FindSV), encapsulating TIDDIT (Eisfeldt et al., 2017) and CNVnator (Abyzov et al., 2011). The subtraction SVs, or somatic SVs, were called using TIDDIT.

2.3.6 Mutation profiles

In *paper IV*, mutations were classified into groups based on known mechanisms of cancer-related genes, which are base substitution, structural rearrangement, copy number variation, and microsatellite instability.

2.3.6.1 Base substitution

Data used for generating base substitution profile were somatic mutation in VCF format. Single base changes were first classified into six subtypes; C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, and T:A > G:C. Then, the changes were further subclassified by including the sequence context of the mutation, which are 5' and 3'. In total, there were 96 mutation types (6 types of substitution x 4 types of 5' base x 4 types of 3' base).

2.3.6.2 Structural rearrangement and copy number variation

Data used in generating SV profiles were the subtraction SV calling performed in an earlier stage. The structural variants were primarily classified into four groups, which are duplication, deletion, inversion, and translocation. Then, they were subclassified based on their approximated size in log10 (size between 100-1Kbp, 1K-10K, 10K-100K, 100K-1M, 1M-10M, 10M-100M, 100M-1000M, and whole chromosome). In total, there were 32 mutation types (4 types of variation x 8 length groups).

2.3.6.3 Microsatellite instability

Primary data used in generating the profile were WGS tumor-normal pairs in BAM format. Detection of MSI loci were called by msisensor (Niu et al., 2014). The mutations were classified based on size- and unique composition of repeat unit (A, C, AC, AG, AT, CG, AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, ATC, CCG, Repeat_unit_length_4, Repeat_unit_length_5).

2.4 DATA ANALYSIS AND VISUALIZATION

2.4.1 Paper I

After data preparation, an analysis of the whole exome was performed (Figure 4). The result variants were then verified with additional 5 members, 2 with rectal cancer, one with 3 TVA, one with 4 TA, and one with 8 HP using sanger sequencing. Then, a segregation in another cohort of 98 familial cancer was performed using the genes with variant found to be segregated in 5 affected members (Figure 5)

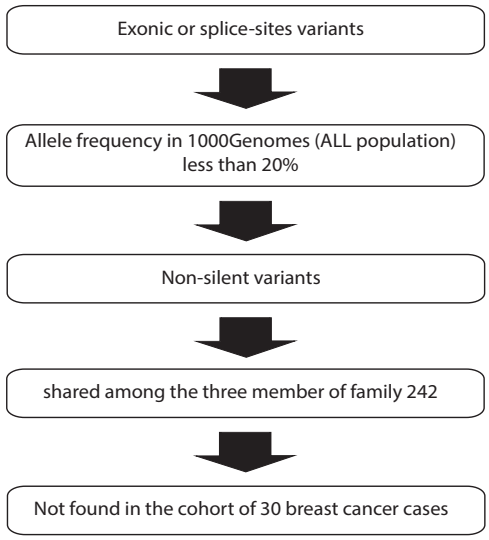

Figure 4. The workflow used for finding variants segregated in the three affected members of family 242.

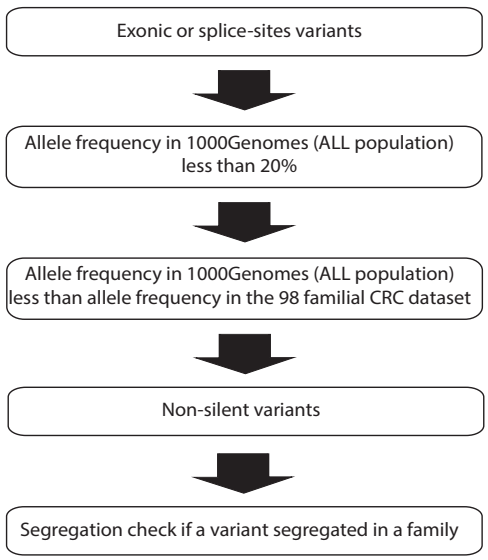

Figure 5. The workflow of segregation analysis in a cohort of 98 familial CRC patients.

2.4.2 Paper II

There were 4 sub studies in this paper. The first one was to look for pathogenic variants in cancer-related genes (Figure 6). The second and the third were to look for possible novel cancer genes using truncating-variant and missense-variant strategies respectively (Figure 7, 8). The fourth was to look for rare monogenic autosomal recessive and less common risk genes (Figure 9) (Thutkawkorapin et al., 2019).

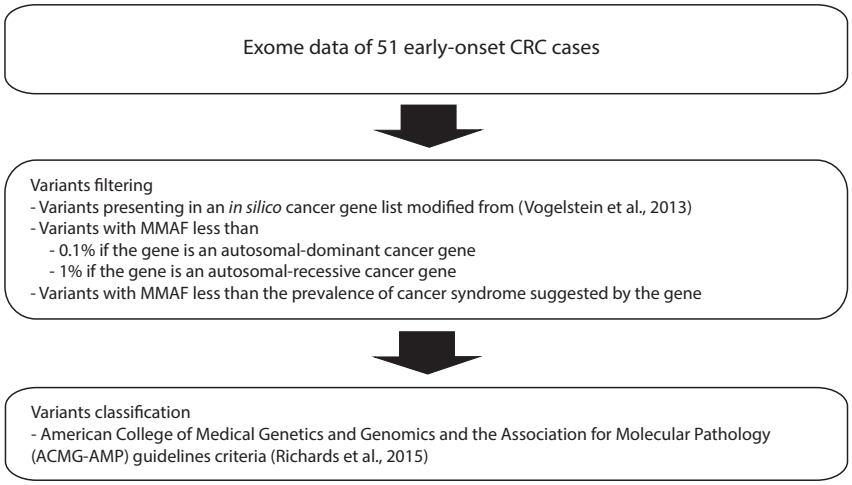

Figure 6. Autosomal dominant and autosomal recessive analysis in cancer susceptibility gene list. [Permission obtained from Wiley to reuse parts of figure 1 from Thutkawkorapin et al. (Thutkawkorapin et al., 2019)]

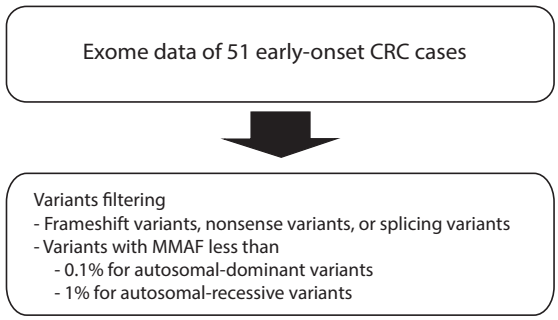

Figure 7. Truncating variant analysis. [Permission obtained from Wiley to reuse parts of figure 2 from Thutkawkorapin et al. (Thutkawkorapin et al., 2019)]

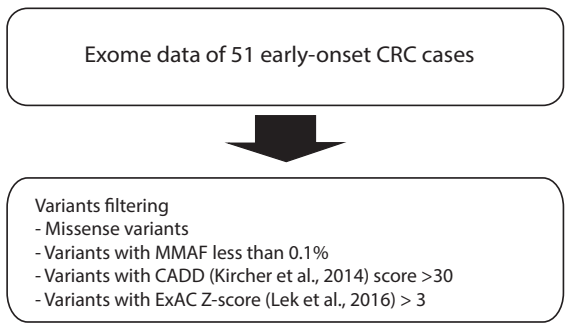

Figure 8. Missense variant analysis. [Permission obtained from Wiley to reuse parts of figure 3 from Thutkawkorapin et al. (Thutkawkorapin et al., 2019)]

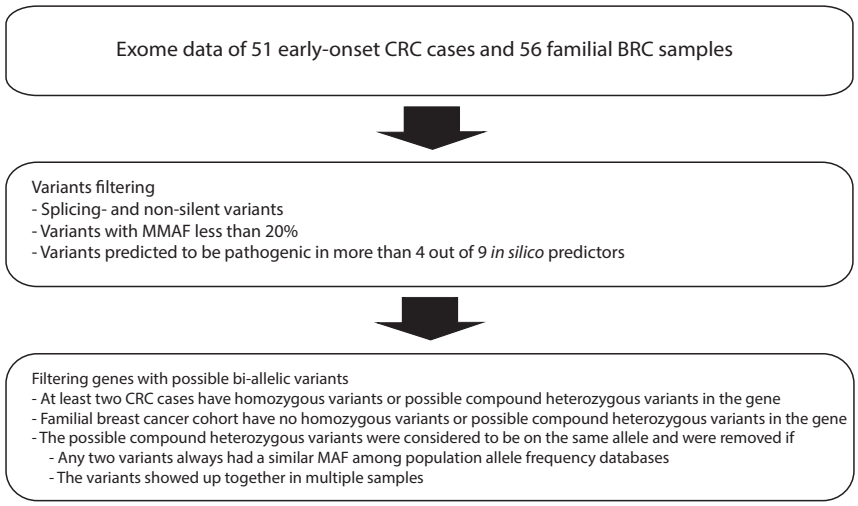

Figure 9. Autosomal recessive gene analysis. [Permission obtained from Wiley to reuse parts of figure 4 from Thutkawkorapin et al. (Thutkawkorapin et al., 2019)]

2.4.3 Paper III

Principal component analysis was performed on SNP array data of samples from STR and NSPHS together with samples from 1000Genomes in order to visualize geographical distribution. After that, the selected 1000 samples were whole genome sequenced. The WGS samples were made publicly available. The allele frequency VCF file was downloadable at https://swefreq.nbis.se/dataset/SweGen. The genome browser of the dataset was developed using software from ExAC (Lek et al., 2016).

2.4.4 Paper IV

The first part of the analysis in this paper was to build up a mutation profile for each sample. the mutation profile consists of a fixed percentage of mutation types from each mutation group. Base substitution accounted for 70% of the profile. Copy number variation (duplication + deletion), structural inversion, and structural rearrangement, together, accounted for 20% of the profile. And the rest 10% of the profile was for microsatellite instability. Within each mutation group, the profile of each mutation type is the relative percentage of the mutation type within the group. For example, if there are 10 events of "C > G" mutation, where both the 5' and 3' bases are A, and the total number of base substitution events are 100, the percentage of this variable would be 70% x 10/100, which would be 7%.

The next step was to find similar patterns underlying the profiles. This was done using a clustering method called non-negative matrix factorization (NMF) to decipher matrix P from given input matrix M, where $M \approx P \times E$. Matrix M represents fraction of each mutation type, the mutation profile, in each sample, each column for one sample and each row for one mutation type. Matrix P represents fraction of mutation type in each cancer process, each column for one cancer signature process and each row for one mutation type. Matrix E

represents fraction of cancer signature process in each sample, each column for one sample and each row for one cancer signature process.

The last step was to visualize the reconstructed profile, matrix newM, when P x E -> newM, together with the original profile and the percentage of the mutational signature components. The purpose was to see the underlying mutational patterns of each samples.

3 RESULTS AND DISCUSSION

3.1 PAPER I

We identified 12 novel non-synonymous single nucleotide variants shared among 5 members with rectal cancer. The mutations were found in 12 different genes; *DZIP1L*, *PCOLCE2*, *IGSF10*, *SUCNR1*, *OR13C8*, *EPB41L4B*, *SEC16A*, *NOTCH1*, *TAS2R7*, *SF3A1*, *GAL3ST1* and *TRIOBP*.

To find a support for as being a high-penetrant gene, we performed a segregation analysis in another cohort of 98 familial colorectal cancer cases (Figure 5). We searched for other mutations in the 12 genes. After this, 36 variants among 11 genes remained. No additional mutation was seen in *SUCNR1*. The result showed that there was a variant in the gene *IGSF10* was shared between two affected relatives in a family. However, the same variant was also found in three other families where it did not segregate with disease. Therefore, none of the genes was suggested as a high penetrant gene.

Considering family members, if we included the member with 3 large tubulovillous adenomas as an affected, three genes, *OR13C8, EPB41L4B* and *TAS2R7,* will be excluded. And if we included the member with had 4 tubular adenomas and 8 hyperplastic polyps, another two genes, *DZIP1L* and *PCOLCE2*, will be excluded. And if we included the member with 5 large hyperplastic polyps, three more genes, *SF3A1*, *GAL3ST1* and *TRIOBP*, will be excluded. We could have used the wrong individuals for our first experiment. In the case one of the three is actually a phenocopy, or if there are two traits, one with high-penetrant gastric cancer and one with high-penetrant rectal cancer, it would have been missed in the analysis. It's also possible that there are two different low-penetrant genes, one for gastric cancer and one for rectal cancer, with the same or different modifying genes among family members.

Based on known functions, *DZIP1L*, *IGSF10*, *NOTCH1*, *SF3A1* and *GAL3ST1,* were proposed to be the candidates. The most likely candidate was *NOTCH1* as it is the best-known gene.

One strong hypothesis in this study was that there was a molecular process involving the risk of developing gastric- and colorectal cancer in this family. The weakness of this hypothesis is that if one, or more, of the 5 affected samples have a phenocopy, causing by different cancer processes, or if the gastric cancer was developed from cancer process different from those of the colorectal cancer, there would be several different combinations to mark the family members as affected. Thus, the study is likely to miss the candidate mutation.

In *paper IV*, we developed a tool to visualize molecular profile of a tumor. The tool will be useful in defining hypothesis for this family to identify samples with the same molecular profile, suggesting the same underlying cancer processes.

3.2 PAPER II

In this study, we used a cohort of 51 CRC cases with an age of onset less than 40 year to search for (1) Rare autosomal dominant and autosomal recessive mutations in candidate cancer susceptibility genes (2) Novel monogenic cancer genes that cause a rare autosomal dominant or autosomal recessive colorectal cancer syndrome in exome data (3) rare monogenic and less common risk genes in exome data.

In (1), we have identified a pathogenic variant in *PTEN*. The patient was subsequently confirmed to have a hereditary hamartoma tumor syndrome. Beside *PTEN*, we found other 7 heterozygous variants in 7 candidate genes. Among them, the variants in *BMPR1A*, *BRIP1* and *SRC*, have never been reported in healthy individuals, making them more likely than the others. In a separate study, immunohistochemistry of mismatch repair genes was performed on one of the patients. His tumor showed loss of *PMS2* protein. After that, target sequencing, using nested PCR, has confirmed the finding with a variant NM_000535.5:c.2113G>A, p.Glu705Lys. The variant was known to be pathogenic and has been reported as causative in many families. The variant in *PMS2* was missed in the study because the sequencing region was not unique resulting in variant with low quality, thus, excluded from the study.

In (2), we used truncating-variants approach and missense-variants approach. In truncating-variants approach, in addition to *PTEN*, the study identified 10 candidate truncating variants in 10 genes. Among this, 7 never been seen in public population. *CLSPN* is the mostly likely based on its function involving DNA damage checkpoint (Chini and Chen, 2003) and DNA damage repair (Azenha et al., 2017). To date, no studies have implicated, SEC24B, or SSH2 in CRC or cancer development. As the variants in SEC24B and SSH2 have never been reported in the normal population databases, they are also possible candidate genes.

In missense-variants approach, using MMAF less than 0.1% alone reduced number of variants down to 3,800 variants. In order to reduce them to a manageable number, we have used ExAC missense z-score more than 3 and CADD score more than 30 as filtering criteria. And finally, we filtered variants that never been seen in public population. These resulted in 8 candidate variants. Considering their functions, *ACACA*, *NR2C2*, *INPP4A*, and *DIDO1* were the proposed candidates. *INPP4A* has been shown to inhibit cell proliferation and promote apoptosis in bladder and pancreatic cancer cells (Wang et al., 2017). Inhibition of *ACACA* can lead to either decreased apoptosis (Keenan et al., 2015) or decreased cell proliferation (Jones et al., 2017, Singh et al., 2015). The inhibition can also lead to increased risk of metastasis/tumor recurrence (in mice) (Rios Garcia et al., 2017). The inhibition of *NR2C2* induces cell death (McNew et al., 2016, Singh et al., 2012). *DIDO1* is upregulated in colorectal tumors (Braig and Bosserhoff, 2013, Sillars-Hardebol et al., 2012). However, this filtering approach was not optimal. List of genes suggested by ExAC z-score to have number of missense mutations less than expected, z-score > 3, will miss candidate genes, for example *POLE* with z-score 1.57. CADD score also showed discrepancies with expert classification from InSiGHT (van der Velde et al., 2015). A recent study, aimed at evaluating the specificity of in silico predictors using common variants (allele frequency > 1% and < 25%) from ExAC

database, showed that the performances vary widely, from 63.4% to 95.5% (Niroula and Vihinen, 2019).

In (3), there were no rare biallelic variants in any known cancer genes found. Then, we searched the whole exome for more common possible biallelic variants, we have listed 16 candidate genes. Among these, 6 of them, *ATP10B*, *PKHD1*, *PTPRQ*, *UGGT2*, *MYH13* and *TFF3*, were more likely based on their MMAF < 5% and observed frequency 20 times higher than the expected likelihood of the two variants occurring together. To search for possible biallelic variants, we used a cohort of 56 breast cancer cases as a comparison group. This may not be a good approach because it can overlook genes predisposing to both colorectal- and breast cancer. However, to search for possible biallelic variants, we need to know the complete genotyping information. One strong advantage of using this inhouse cohort is in its ability to remove platform errors. The breast-cancer cohort is from Swedish population, the same as the CRC. Moreover, the DNA from both cohorts has been collected and processed in the same way. Thus, it can eliminate the artifacts caused by the differences in populations and the difference in platforms

3.3 PAPER III

The goal of this study was to provide a local reference database of 1,000 Swedish individuals to the public. DNA samples in this study were from the Swedish Twin Registry (STR) (n=10,000) and Northern Sweden Population Health Study (NSPHS) (n=1,033). In order to select samples that can represent Swedish population geographically, this study has performed principal component analysis (PCA) using SNP array data of STR and NSPHS samples together with SNVs extracted from 1000 Genomes project phase 3 (Figure 9). The result showed that genetic structure and STR and NSPHS is likely to represent Swedish individuals who have been living in Sweden for at least one generation. Using PCA, 942 STR samples and 58 NSPHS samples were selected for whole genome sequencing.

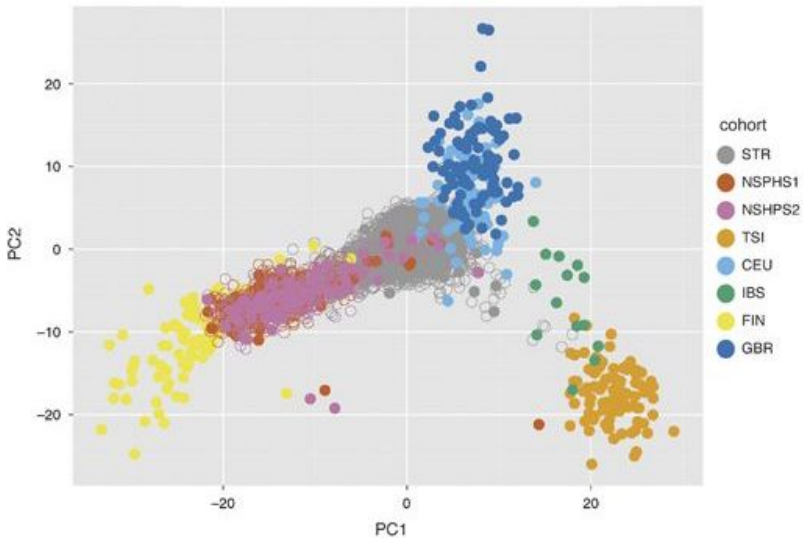

Figure 10. Principal analysis of SNP array data. STR: the Swedish Twin Registry, NSPHS1 and NSPHS2: the Northern Sweden Population Health Study (collected in two different phases), CEU: Utah Residents with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Toscani in Italia. [Permission obtained from SpringerNature to reuse parts of figure 1 from Ameur et al. (Ameur et al., 2017)]

In order to evaluate this whole genome sequencing dataset, this study has performed an overview of SNVs and indels. This study revealed 8.9 million SNVs and 1.0 million indels not presenting in dbSNP (version 147). Of this, 23,396 SNVs and 3239 indels were found to have consequence in amino acid change. This indicated that a large number of genetic variants not currently reported in dbSNP of Swedish population could have a direct effect on protein.

In order to evaluate the WGS data in relation to other populations, this study has performed another PCA using SNVs extracted from the WGS data, excluding 58 NSPHS samples due to their higher degree of relatedness, and 1000 genomes phase 3 (Figure 11). The analysis showed that the STR cohort is genetically close to the European population in 1000 Genomes (Figure

11a). Among European samples, the study showed that the STR samples are intermixed with the Finnish samples (Figure 11b). This supported the finding of a high degree of genetic diversity of the selected Swedish samples.

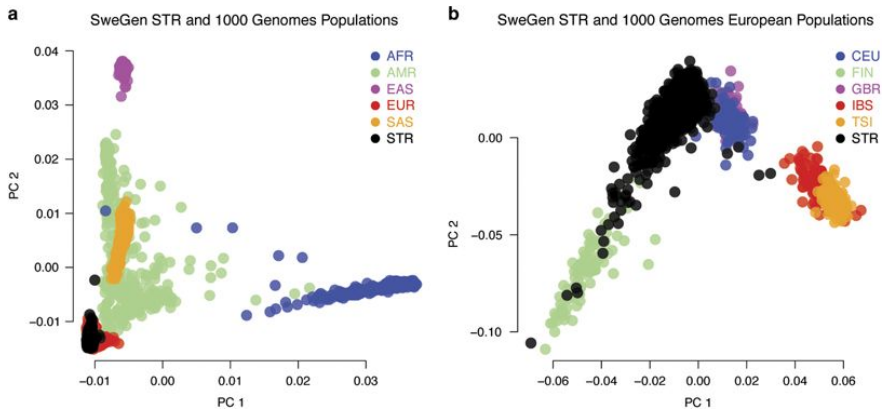

Figure 11. Genetic variation of the 942 Swegen STR samples in relation to 1000 Genomes populations. (a) PCA result in comparison with 1000 Genomes global sub-populations (AFR=African, AMR=Ad Mixed American, EAS=East Asian, EUR=European, SAS=South Asian). (b) PCA result in comparison with 1000 Genomes European sub-populations (CEU: Utah Residents with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Toscani in Italia). [Permission obtained from SpringerNature to reuse parts of figure 1 from Ameur et al. (Ameur et al., 2017)]

This study has made the variant frequency database of 1,000 Swedish individuals available at https://swefreq.nbis.se/dataset/SweGen/browser. The genome browser was developed using software from ExAC. My contribution in this project was to incorporate genetic variation found in this study into the browser platform, as it was not optimized for whole genome sequencing data.

This local reference database of Swedish population has proved to be very useful. To date, it has been cited by 30 articles, including the *paper II* in this thesis. As demonstrated in the study, there is a large proportion of genetic variation found in this study unique to Swedish population. Including this database as another filtering criteria, as done in paper II, will help removing local common variants, otherwise interpreted as rare. Interestingly, a recent study showed that some of in silico predictors have poor specificity in predicting benign effect of variants common in one population but unique in all other population (AF > 1% but < 25%) (Niroula and Vihinen, 2019). They tended to annotate the variants as disease-causing or benign based on dbSNP, 1000Genomes and ClinVar. This local database will exclude the false positive bias reported by these tools.

3.4 PAPER IV

The aim of this study is to use an unsupervised learning approach to identify combinations of both known and unknown cancer processes in humans based on the integration of base substitution profile, copy number variation profile, structural rearrangement profile and microsatellite instability profile in 74 whole genome sequencing tumor-normal pairs from The Cancer Genome Atlas project (TCGA). The cohort consists of 68 breast- and 8 colorectal cancer cases.

The results identified 12 patterns, both known as well as novel, of underlying cancer processes and suggested associations between different types of DNA damage. 9 signatures were unique to the breast cancer cohort and 2 signatures were unique to colorectal cancer cohort (Figure 5). There was only one process associated with both cancer type suggesting the mutational processes in cancer were not completely random.

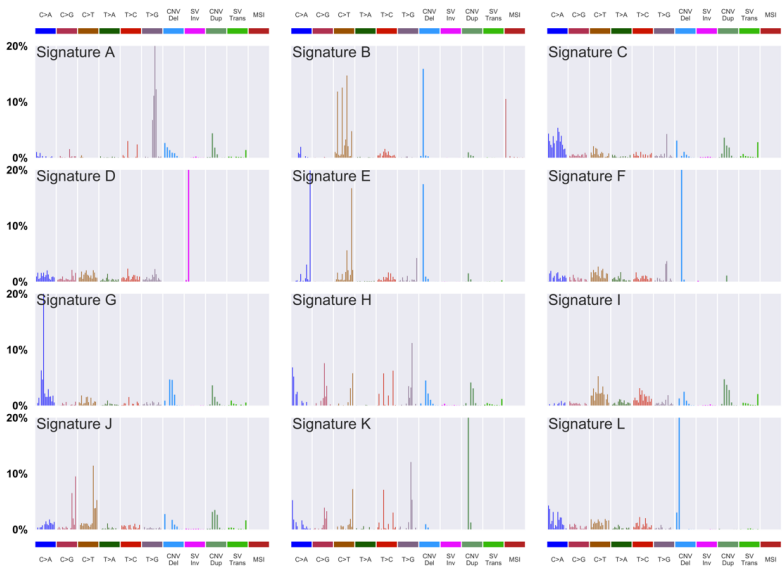

Figure 5. 12 profile patterns found in 74 samples from TCGA. Signature A, C, D, F, G, H J, K and L were associated only with breast cancer cases while Signature B and E were associated with only colorectal cancer cases. Only Signature I was associated with both cancer types.

Signature B was associated with two colorectal cancer samples suggested by msisensor (Niu et al., 2014) to have microsatellite instability. Signature E was associated with two colorectal cancer samples with somatic mutation in *POLE*. For breast cancer samples, 3 out of 4 samples with somatic mutation in *BRCA1* have similar structural variation profiles and have the same biggest 2 signatures.

In samples with high SV burden, which was described as having high number of structural events, the highest 10 breast cancer samples were associated with variety of signatures, C(n=1),

D(n=4), F(n=4), and H(n=1). However, if we looked at only the structural variation profiles, they shared one similarity. Their SV profiles were dominated by: deletion events with size between 10K – 100K; inversion events with size between 1K – 10K; and duplication events with size between 10K – 100K. In other words, total amount of these 3 mutation events always represented more than 80% of their SV profile.

The conclusion from the study so far is that including multiple variant types allowed the model to identify underlying patterns connecting these data types. It suggested that there was a connection between mutation types as in *POLE* and MSI cases. But that was not always correct, as seen in SV burden cases, in which the unclear pattern in base substitution profile may have diluted the performance of pattern finding. Currently, the profile extraction method has a bias toward base substitution, 70% of the profile weight. This is because there are 96 mutation types for base substitution compared to 32 mutation types of SV and 14 of MSI. The result from this study shares similar conclusions with a recent study by (Funnell et al., 2019) which emphasize that integrating multiple variant types can reveal corelated structures both within and between the data types and increase the accuracy of discovering mutational patterns.

The original intent of this study was to analyse all WGS tumor-normal pairs from TCGA (n=503) with an aim to discover underlying mutational patterns, similar to https://cancer.sanger.ac.uk/cosmic/signatures, by integrating multiple variant types. However, the total data size is really big and it take very long time for downloading the data. The study done here served as a feasibility study to prove that integrating more variant types can reveal more mutational structures and reveal more association between mutation types.

The strength of unsupervised methods is in its un-bias toward any predefined hypothesis (e.g. a monogenic disease, a gene-related disease, a pathway-related disease). Its aim is to find similarity between samples. However, the result from this may not be usable right away, as there is no sample labelling in the input. Additional steps are required in order to make the result usable. Etiological study for each sample is needed to truly confirm if the sample is caused by a monogenic disease, or a pathway-related disease. The results from the etiological study will label the sample. The pattern identified by this model and the correct labelling can be further used in a supervised learning model to develop a predictor that can predict the correlation between the mutational profile and the clinical phenotype or to predict the correlation between the mutational profile and the clinical outcome

4 ACKNOWLEDGEMENTS

For the most important part of my study, I would like to thank all the **patients and their families**, together with **every staffs** involved.

To my former main supervisor, **Annika Lindblom**, for giving me the opportunity to join her group, for introducing me to cancer study, for your guidance, for your support, and, most importantly, for the chance to develop into a scientist.

To my current main supervisor, **Emma Tham**, for your careful attention, for your time, for your detailed explanation, and for your close supervised. I really like it when you asked "What will be my biggest problem in this defence?"

To my bioinformatics supervisor, **Daniel Nilsson**, for allowing me to do my master thesis, for introducing me to Annika, for scientific ideas, for very nice discussions, for always looking at the big picture, and for freedom of thinking.

To current- and former members in cancer genetics group, **Patrick Bryant**, **Xiang Jiao**, **Wen Liu**, **Hovsep Mahdessian**, **Tao Liu**, **Susanna von Holst**, **Vinay kumar Kontham**, **Anna-Lena Kastman**, **Johanna Rantala**, and **people in the clinics**.

To nearby PhD students, who are sharing the same fate and help me with miscellaneous issues, giving me a nice feeling that I'm not the only one with this interesting experience, **Alice Costantini**, **Maria Pettersson**, **Anna Hammarsjö**, and **J35P312 Eisfeldt**.

To **everyone on J10:20**. It's nice to have lunch, fika, mingles, and discussed random stuffs with you from time to time. Also for the helps in many occasions. Special thanks to **Nina Jäntt** and **Raquel Vaz,** your "how are you?" is not just greeting.

To **friends at CMM and MTC**, special thanks to **Sherwin**, **SiSi**, **Pilar**, **Atticus**, **Paola**, **Hodan**, **Daisy**, and **Zul**.

To **researchers**, **clinicians**, and **co-authors** that have contributed to the projects in this thesis, thanks for your hard work.

To administrative staffs at MMK, special thanks to **Ann-Britt Wikström.**

To **Ja-aye**, for our very long friendship and for introducing Sweden to me.

To **all of my friends**, who have supported me all this time.

To my family, **my dad**, **my aunt**, **Lonk**, **Eve**, for your support. Special thanks to **Ping**, without you over there, I wouldn't be here.

To **my mom** and **my grandma**, for your forever support.

To **Jeab**, for becoming the most important part of my life.

5 REFERENCES

Cancer Research UK [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer/incidence#heading-One [Accessed April 11 2016].

The Genome Reference Consortium [Online]. Available: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/ [Accessed].

AARTSMA-RUS, A., VAN DEUTEKOM, J. C., FOKKEMA, I. F., VAN OMMEN, G. J. & DEN DUNNEN, J. T. 2006. Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle Nerve,* 34**,** 135-44.

ABBS, S., BUSSOLI, T. & KAVALIER, F. 2004. Nature Encyclopaedia of the Human Genome. *BMJ : British Medical Journal,* 328**,** 172-172.

ABYZOV, A., URBAN, A. E., SNYDER, M. & GERSTEIN, M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res,* 21**,** 974-84.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AGNOLI, C., GRIONI, S., SIERI, S., PALLI, D., MASALA, G., SACERDOTE, C., VINEIS, P., TUMINO, R., GIURDANELLA, M. C., PALA, V., BERRINO, F., MATTIELLO, A., PANICO, S. & KROGH, V. 2013. Italian Mediterranean Index and risk of colorectal cancer in the Italian section of the EPIC cohort. *Int J Cancer,* 132**,** 1404-11.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., CAMPBELL, P. J. & STRATTON, M. R. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep,* 3**,** 246-59.

ALM, T. 1975. Surgical treatment of hereditary adenomatosis of the colon and rectum in Sweden during the last 20 years. Part II. Patients with prophylactic operations, primary and late results. Discussion and summary. *Acta Chir Scand,* 141**,** 228-37.

AMBROS, V. 2004. The functions of animal microRNAs. *Nature,* 431**,** 350-5.

AMEUR, A., DAHLBERG, J., OLASON, P., VEZZI, F., KARLSSON, R., MARTIN, M., VIKLUND, J., KAHARI, A. K., LUNDIN, P., CHE, H., THUTKAWKORAPIN, J., EISFELDT, J., LAMPA, S., DAHLBERG, M., HAGBERG, J., JAREBORG, N., LILJEDAHL, U., JONASSON, I., JOHANSSON, A., FEUK, L., LUNDEBERG, J., SYVANEN, A. C., LUNDIN, S., NILSSON, D., NYSTEDT, B., MAGNUSSON, P. K. & GYLLENSTEN, U. 2017. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet,* 25**,** 1253-1260.

ARNOLD, S., BUCHANAN, D. D., BARKER, M., JASKOWSKI, L., WALSH, M. D., BIRNEY, G., WOODS, M. O., HOPPER, J. L., JENKINS, M. A., BROWN, M. A., TAVTIGIAN, S. V., GOLDGAR, D. E., YOUNG, J. P. & SPURDLE, A. B. 2009. Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum Mutat,* 30**,** 757-70.

AZENHA, D., LOPES, M. C. & MARTINS, T. C. 2017. Claspin functions in cell homeostasis-A link to cancer? *DNA Repair (Amst),* 59**,** 27-33.

BANDRES, E., CUBEDO, E., AGIRRE, X., MALUMBRES, R., ZARATE, R., RAMIREZ, N., ABAJO, A., NAVARRO, A., MORENO, I., MONZO, M. & GARCIA-FONCILLAS, J. 2006. Identification by Real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues. *Mol Cancer,* 5**,** 29.

BAO, R., HERNANDEZ, K., HUANG, L., KANG, W., BARTOM, E., ONEL, K., VOLCHENBOUM, S. & ANDRADE, J. 2015. ExScalibur: A High-Performance Cloud-Enabled Suite for Whole Exome Germline and Somatic Mutation Identification. *PLoS One,* 10**,** e0135800.

BARTEL, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-97.

BLACKWOOD, E. M. & KADONAGA, J. T. 1998. Going the distance: a current view of enhancer action. *Science,* 281**,** 60-3.

BOBADILLA, J. L., MACEK, M., JR., FINE, J. P. & FARRELL, P. M. 2002. Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Hum Mutat,* 19**,** 575-606.

BODMER, W. F., BAILEY, C. J., BODMER, J., BUSSEY, H. J., ELLIS, A., GORMAN, P., LUCIBELLO, F. C., MURDAY, V. A., RIDER, S. H., SCAMBLER, P. & ET AL. 1987. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature,* 328**,** 614-6.

BOLAND, C. R., THIBODEAU, S. N., HAMILTON, S. R., SIDRANSKY, D., ESHLEMAN, J. R., BURT, R. W., MELTZER, S. J., RODRIGUEZ-BIGAS, M. A., FODDE, R., RANZANI, G. N. & SRIVASTAVA, S. 1998. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res,* 58**,** 5248-57.

BRAIG, S. & BOSSERHOFF, A. K. 2013. Death inducer-obliterator 1 (Dido1) is a BMP target gene and promotes BMP-induced melanoma progression. *Oncogene,* 32**,** 837-48.

BURT, R. 2007. Inheritance of Colorectal Cancer. *Drug Discov Today Dis Mech,* 4**,** 293-300.

BURT, R. W. 2000. Colon cancer screening. *Gastroenterology,* 119**,** 837-53.

BYARD, P. J. 1994. The adolescent growth spurt in children with cystic fibrosis. *Ann Hum Biol,* 21**,** 229-40.

CANCER GENOME ATLAS, N. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-7.

CANTAREL, B. L., WEAVER, D., MCNEILL, N., ZHANG, J., MACKEY, A. J. & REESE, J. 2014. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics,* 15**,** 104.

CAVENEE, W. K., DRYJA, T. P., PHILLIPS, R. A., BENEDICT, W. F., GODBOUT, R., GALLIE, B. L., MURPHREE, A. L., STRONG, L. C. & WHITE, R. L. 1983. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature,* 305**,** 779-84.

CHEN, X., SCHULZ-TRIEGLAFF, O., SHAW, R., BARNES, B., SCHLESINGER, F., KALLBERG, M., COX, A. J., KRUGLYAK, S. & SAUNDERS, C. T. 2015. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*.

CHINI, C. C. & CHEN, J. 2003. Human claspin is required for replication checkpoint control. *J Biol Chem,* 278**,** 30057-62.

CHUBB, D., BRODERICK, P., FRAMPTON, M., KINNERSLEY, B., SHERBORNE, A., PENEGAR, S., LLOYD, A., MA, Y. P., DOBBINS, S. E. & HOULSTON, R. S. 2015. Genetic diagnosis of high-penetrance susceptibility for colorectal cancer (CRC) is achievable for a high proportion of familial CRC by exome sequencing. *J Clin Oncol,* 33**,** 426-32.

CHUN, S. & FAY, J. C. 2009. Identification of deleterious mutations within three human genomes. *Genome Res,* 19**,** 1553-61.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

COOPER, G. M., STONE, E. A., ASIMENOS, G., PROGRAM, N. C. S., GREEN, E. D., BATZOGLOU, S. & SIDOW, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res,* 15**,** 901-13.

DAVIES, H., GLODZIK, D., MORGANELLA, S., YATES, L. R., STAAF, J., ZOU, X., RAMAKRISHNA, M., MARTIN, S., BOYAULT, S., SIEUWERTS, A. M., SIMPSON, P. T., KING, T. A., RAINE, K., EYFJORD, J. E., KONG, G., BORG, A., BIRNEY, E., STUNNENBERG, H. G., VAN DE VIJVER, M. J., BORRESEN-DALE, A. L., MARTENS, J. W., SPAN, P. N., LAKHANI, S. R., VINCENT-SALOMON, A., SOTIRIOU, C., TUTT, A., THOMPSON, A. M., VAN LAERE, S., RICHARDSON, A. L., VIARI, A., CAMPBELL, P. J., STRATTON, M. R. & NIK-ZAINAL, S. 2017. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med,* 23**,** 517-525.

DAVYDOV, E. V., GOODE, D. L., SIROTA, M., COOPER, G. M., SIDOW, A. & BATZOGLOU, S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol,* 6**,** e1001025.

DE MAGALHAES, J. P., FINCH, C. E. & JANSSENS, G. 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res Rev,* 9**,** 315-23.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.

DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., XUE, C., MARINOV, G. K., KHATUN, J., WILLIAMS, B. A., ZALESKI, C., ROZOWSKY, J., RODER, M., KOKOCINSKI, F., ABDELHAMID, R. F., ALIOTO, T., ANTOSHECHKIN, I., BAER, M. T., BAR, N. S., BATUT, P., BELL, K., BELL, I., CHAKRABORTTY, S., CHEN, X., CHRAST, J., CURADO, J., DERRIEN, T., DRENKOW, J., DUMAIS, E., DUMAIS, J., DUTTAGUPTA, R., FALCONNET, E., FASTUCA, M., FEJES-TOTH, K., FERREIRA, P., FOISSAC, S., FULLWOOD, M. J., GAO, H., GONZALEZ, D., GORDON, A., GUNAWARDENA, H., HOWALD, C., JHA, S., JOHNSON, R., KAPRANOV, P., KING, B., KINGSWOOD, C., LUO, O. J., PARK, E., PERSAUD, K., PREALL, J. B., RIBECA, P., RISK, B., ROBYR, D., SAMMETH, M., SCHAFFER, L., SEE, L. H., SHAHAB, A., SKANCKE, J., SUZUKI, A. M., TAKAHASHI, H., TILGNER, H., TROUT, D., WALTERS, N., WANG, H., WROBEL, J., YU, Y., RUAN, X., HAYASHIZAKI, Y., HARROW, J., GERSTEIN, M., HUBBARD, T., REYMOND, A., ANTONARAKIS, S. E., HANNON, G., GIDDINGS, M. C., RUAN, Y., WOLD, B., CARNINCI, P., GUIGO, R. & GINGERAS, T. R. 2012. Landscape of transcription in human cells. *Nature,* 489**,** 101-8.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res,* 36**,** e105.

EGGER, G., LIANG, G., APARICIO, A. & JONES, P. A. 2004. Epigenetics in human disease and prospects for epigenetic therapy. *Nature,* 429**,** 457-63.

EISFELDT, J., VEZZI, F., OLASON, P., NILSSON, D. & LINDSTRAND, A. 2017. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res,* 6**,** 664.

ERIK GARRISON, G. M. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint*.

ESTEBAN-JURADO, C., VILA-CASADESUS, M., GARRE, P., LOZANO, J. J., PRISTOUPILOVA, A., BELTRAN, S., MUNOZ, J., OCANA, T., BALAGUER, F., LOPEZ-CERON, M., CUATRECASAS, M., FRANCH-EXPOSITO, S., PIQUE, J. M., CASTELLS, A., CARRACEDO, A., RUIZ-PONTE, C., ABULI, A., BESSA, X., ANDREU, M., BUJANDA, L., CALDES, T. & CASTELLVI-BEL, S. 2015. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet Med,* 17**,** 131-42.

FANG, H., WU, Y., NARZISI, G., O'RAWE, J. A., BARRON, L. T., ROSENBAUM, J., RONEMUS, M., IOSSIFOV, I., SCHATZ, M. C. & LYON, G. J. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med,* 6**,** 89.

FARRINGTON, S. M., TENESA, A., BARNETSON, R., WILTSHIRE, A., PRENDERGAST, J., PORTEOUS, M., CAMPBELL, H. & DUNLOP, M. G. 2005. Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am J Hum Genet,* 77**,** 112-9.

FEARON, E. R. & VOGELSTEIN, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61.

FODDE, R., KUIPERS, J., ROSENBERG, C., SMITS, R., KIELMAN, M., GASPAR, C., VAN ES, J. H., BREUKEL, C., WIEGANT, J., GILES, R. H. & CLEVERS, H. 2001.

Mutations in the APC tumour suppressor gene cause chromosomal instability. *Nat Cell Biol,* 3**,** 433-8.

FORBES, S. A., BEARE, D., BOUTSELAKIS, H., BAMFORD, S., BINDAL, N., TATE, J., COLE, C. G., WARD, S., DAWSON, E., PONTING, L., STEFANCSIK, R., HARSHA, B., KOK, C. Y., JIA, M., JUBB, H., SONDKA, Z., THOMPSON, S., DE, T. & CAMPBELL, P. J. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res,* 45**,** D777-D783.

FRATKIN, E., BERCOVICI, S. & STEPHAN, D. A. 2012. The implications of ENCODE for diagnostics. *Nat Biotechnol,* 30**,** 1064-5.

FROMER, M., MORAN, J. L., CHAMBERT, K., BANKS, E., BERGEN, S. E., RUDERFER, D. M., HANDSAKER, R. E., MCCARROLL, S. A., O'DONOVAN, M. C., OWEN, M. J., KIROV, G., SULLIVAN, P. F., HULTMAN, C. M., SKLAR, P. & PURCELL, S. M. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet,* 91**,** 597-607.

FUNNELL, T., ZHANG, A. W., GREWAL, D., MCKINNEY, S., BASHASHATI, A., WANG, Y. K. & SHAH, S. P. 2019. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput Biol,* 15**,** e1006799.

GALINSKY, V. L. 2012. YOABS: yet other aligner of biological sequences--an efficient linearly scaling nucleotide aligner. *Bioinformatics,* 28**,** 1070-7.

GISMONDI, V., META, M., BONELLI, L., RADICE, P., SALA, P., BERTARIO, L., VIEL, A., FORNASARIG, M., ARRIGONI, A., GENTILE, M., PONZ DE LEON, M., ANSELMI, L., MARENI, C., BRUZZI, P. & VARESCO, L. 2004. Prevalence of the Y165C, G382D and 1395delGGA germline mutations of the MYH gene in Italian patients with adenomatous polyposis coli and colorectal adenomas. *Int J Cancer,* 109**,** 680-4.

GOLDSCHMIDT, R. B. 1949. Phenocopies. *Sci Am,* 181**,** 46-9.

GRODY, W. W., THOMPSON, B. H. & HUDGINS, L. 2013. Whole-exome/genome sequencing and genomics. *Pediatrics,* 132**,** S211-5.

GROVER, S., KASTRINOS, F., STEYERBERG, E. W., COOK, E. F., DEWANWALA, A., BURBIDGE, L. A., WENSTRUP, R. J. & SYNGAL, S. 2012. Prevalence and phenotypes of APC and MUTYH mutations in patients with multiple colorectal adenomas. *Jama,* 308**,** 485-92.

HAINAUT, P. & PFEIFER, G. P. 2001. Patterns of p53 G-->T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis,* 22**,** 367-74.

HALF, E., BERCOVICH, D. & ROZEN, P. 2009. Familial adenomatous polyposis. *Orphanet J Rare Dis,* 4**,** 22.

HAMOSH, A., SCOTT, A. F., AMBERGER, J. S., BOCCHINI, C. A. & MCKUSICK, V. A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res,* 33**,** D514-7.

JAGADEESH, K. A., PAGGI, J. M., YE, J. S., STENSON, P. D., COOPER, D. N., BERNSTEIN, J. A. & BEJERANO, G. 2019. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet,* 51**,** 755-763.

JOHNS, L. E. & HOULSTON, R. S. 2001. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol,* 96**,** 2992-3003.

JONES, J. E., ESLER, W. P., PATEL, R., LANBA, A., VERA, N. B., PFEFFERKORN, J. A. & VERNOCHET, C. 2017. Inhibition of Acetyl-CoA Carboxylase 1 (ACC1) and 2 (ACC2) Reduces Proliferation and De Novo Lipogenesis of EGFRvIII Human Glioblastoma Cells. *PLoS One,* 12**,** e0169566.

KASSAHN, K. S., SCOTT, H. S. & CARAMINS, M. C. 2014. Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. *Hum Mutat,* 35**,** 413-23.

KAWAJI, H., SEVERIN, J., LIZIO, M., FORREST, A. R., VAN NIMWEGEN, E., REHLI, M., SCHRODER, K., IRVINE, K., SUZUKI, H., CARNINCI, P., HAYASHIZAKI, Y. & DAUB, C. O. 2011. Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res,* 39**,** D856-60.

KAWAJI, H., SEVERIN, J., LIZIO, M., WATERHOUSE, A., KATAYAMA, S., IRVINE, K. M., HUME, D. A., FORREST, A. R., SUZUKI, H., CARNINCI, P., HAYASHIZAKI, Y. & DAUB, C. O. 2009. The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol,* 10**,** R40.

KEENAN, M. M., LIU, B., TANG, X., WU, J., CYR, D., STEVENS, R. D., ILKAYEVA, O., HUANG, Z., TOLLINI, L. A., MURPHY, S. K., LUCAS, J., MUOIO, D. M., KIM, S. Y. & CHI, J. T. 2015. ACLY and ACC1 Regulate Hypoxia-Induced Apoptosis by Modulating ETV4 via alpha-ketoglutarate. *PLoS Genet,* 11**,** e1005599.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet,* 46**,** 310-5.

KLEMKE, M., DRIESCHNER, N., LAABS, A., RIPPE, V., BELGE, G., BULLERDIEK, J. & SENDT, W. 2011. On the prevalence of the PAX8-PPARG fusion resulting from the chromosomal translocation t(2;3)(q13;p25) in adenomas of the thyroid. *Cancer Genet,* 204**,** 334-9.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

KOHLMANN, W. & GRUBER, S. B. 1993. Lynch Syndrome. *In:* PAGON, R. A., ADAM, M. P., ARDINGER, H. H., WALLACE, S. E., AMEMIYA, A., BEAN, L. J. H., BIRD, T. D., FONG, C. T., MEFFORD, H. C., SMITH, R. J. H. & STEPHENS, K. (eds.) *GeneReviews(R)*. Seattle (WA): University of Washington, Seattle

KONNO, P. & SILM, H. 2001. Waardenburg syndrome. *J Eur Acad Dermatol Venereol,* 15**,** 330-3.

KUMAR, P., HENIKOFF, S. & NG, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4**,** 1073-81.

LAI, Z., MARKOVETS, A., AHDESMAKI, M., CHAPMAN, B., HOFMANN, O., MCEWEN, R., JOHNSON, J., DOUGHERTY, B., BARRETT, J. C. & DRY, J. R. 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*.

LANDRUM, M. J., LEE, J. M., BENSON, M., BROWN, G. R., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., JANG, W., KARAPETYAN, K., KATZ, K., LIU, C., MADDIPATLA, Z., MALHEIRO, A., MCDANIEL, K., OVETSKY, M., RILEY, G., ZHOU, G., HOLMES, J. B., KATTMAN, B. L. & MAGLOTT, D. R. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res,* 46**,** D1062-D1067.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.

LAPPALAINEN, I., LOPEZ, J., SKIPPER, L., HEFFERON, T., SPALDING, J. D., GARNER, J., CHEN, C., MAGUIRE, M., CORBETT, M., ZHOU, G., PASCHALL, J., ANANIEV, V., FLICEK, P. & CHURCH, D. M. 2013. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res,* 41**,** D936-41.

LE MARCHAND, L., WILKENS, L. R., HANKIN, J. H., KOLONEL, L. N. & LYU, L. C. 1997. A case-control study of diet and colorectal cancer in a multiethnic population in Hawaii (United States): lipids and foods of animal origin. *Cancer Causes Control,* 8**,** 637-48.

LEARY, R. J., LIN, J. C., CUMMINS, J., BOCA, S., WOOD, L. D., PARSONS, D. W., JONES, S., SJOBLOM, T., PARK, B. H., PARSONS, R., WILLIS, J., DAWSON, D., WILLSON, J. K., NIKOLSKAYA, T., NIKOLSKY, Y., KOPELOVICH, L., PAPADOPOULOS, N., PENNACCHIO, L. A., WANG, T. L., MARKOWITZ, S. D., PARMIGIANI, G., KINZLER, K. W., VOGELSTEIN, B. & VELCULESCU, V. E. 2008. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A,* 105**,** 16224-9.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H. H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION,

C. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

LETUNIC, I., DOERKS, T. & BORK, P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res,* 40**,** D302-5.

LI, H. 2015. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics,* 31**,** 3694-6.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, H. & DURBIN, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26**,** 589-95.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, J. Y., GAILLARD, F., MOREAU, A., HAROUSSEAU, J. L., LABOISSE, C., MILPIED, N., BATAILLE, R. & AVET-LOISEAU, H. 1999. Detection of translocation t(11;14)(q13;q32) in mantle cell lymphoma by fluorescence in situ hybridization. *Am J Pathol,* 154**,** 1449-52.

LI, R., YU, C., LI, Y., LAM, T. W., YIU, S. M., KRISTIANSEN, K. & WANG, J. 2009b. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-7.

LI, W. D., LI, Q. R., XU, S. N., WEI, F. J., YE, Z. J., CHENG, J. K. & CHEN, J. P. 2013. Exome sequencing identifies an MLL3 gene germ line mutation in a pedigree of colorectal cancer and acute myeloid leukemia. *Blood,* 121**,** 1478-9.

LIEBER, M. R. 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. *Annu Rev Biochem,* 79**,** 181-211.

LIN, J. K., CHANG, S. C., YANG, Y. C. & LI, A. F. 2003. Loss of heterozygosity and DNA aneuploidy in colorectal adenocarcinoma. *Ann Surg Oncol,* 10**,** 1086-94.

LINDBLOM, A., TANNERGARD, P., WERELIUS, B. & NORDENSKJOLD, M. 1993. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nat Genet,* 5**,** 279-82.

LIU, Y. & SCHMIDT, B. 2012. Long read alignment based on maximal exact match seeds. *Bioinformatics,* 28**,** i318-i324.

LOPEZ-OTIN, C., BLASCO, M. A., PARTRIDGE, L., SERRANO, M. & KROEMER, G. 2013. The hallmarks of aging. *Cell,* 153**,** 1194-217.

LUBBE, S. J., DI BERNARDO, M. C., CHANDLER, I. P. & HOULSTON, R. S. 2009. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol,* 27**,** 3975-80.

LUNTER, G. & GOODSON, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res,* 21**,** 936-9.

LUO, R., WONG, T., ZHU, J., LIU, C. M., ZHU, X., WU, E., LEE, L. K., LIN, H., ZHU, W., CHEUNG, D. W., TING, H. F., YIU, S. M., PENG, S., YU, C., LI, Y., LI, R. & LAM, T. W. 2013. SOAP3-dp: fast, accurate and sensitive GPU-based short read aligner. *PLoS One,* 8**,** e65632.

MAILMAN, M. D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., ZIYABARI, L., LEE, M., SHAO, Y., WANG, Z. Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J. & SHERRY, S. T. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet,* 39**,** 1181-6.

MALT, E. A., DAHL, R. C., HAUGSAND, T. M., ULVESTAD, I. H., EMILSEN, N. M., HANSEN, B., CARDENAS, Y. E., SKOLD, R. O., THORSEN, A. T. & DAVIDSEN, E. M. 2013. Health and disease in adults with Down syndrome. *Tidsskr Nor Laegeforen,* 133**,** 290-4.

MARDIS, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet,* 24**,** 133-41.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol,* 17**,** 122.

MCNEW, K. L., WHIPPLE, W. J., MEHTA, A. K., GRANT, T. J., RAY, L., KENNY, C. & SINGH, A. 2016. MEK and TAK1 Regulate Apoptosis in Colon Cancer Cells with KRAS-Dependent Activation of Proinflammatory Signaling. *Mol Cancer Res,* 14**,** 1204-1216.

MEYNERT, A. M., ANSARI, M., FITZPATRICK, D. R. & TAYLOR, M. S. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics,* 15**,** 1-11.

MOLLER, P., SEPPALA, T., BERNSTEIN, I., HOLINSKI-FEDER, E., SALA, P., EVANS, D. G., LINDBLOM, A., MACRAE, F., BLANCO, I., SIJMONS, R., JEFFRIES, J., VASEN, H., BURN, J., NAKKEN, S., HOVIG, E., RODLAND, E. A., THARMARATNAM, K., DE VOS TOT NEDERVEEN CAPPEL, W. H., HILL, J., WIJNEN, J., GREEN, K., LALLOO, F., SUNDE, L., MINTS, M., BERTARIO, L., PINEDA, M., NAVARRO, M., MORAK, M., RENKONEN-SINISALO, L., FRAYLING, I. M., PLAZZER, J. P., PYLVANAINEN, K., SAMPSON, J. R., CAPELLA, G., MECKLIN, J. P. & MOSLEIN, G. 2015. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut*.

NIELSEN, M., LYNCH, H., INFANTE, E. & AL, E. 2012. *MUTYH-Associated Polyposis* [Online]. Seattle (WA): University of Washington, Seattle: GeneReviews. Available: http://www.ncbi.nlm.nih.gov/books/NBK107219/ [Accessed April 11 2016].

NIELSEN, M., MORREAU, H., VASEN, H. F. & HES, F. J. 2011. MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol,* 79**,** 1-16.

NIROULA, A. & VIHINEN, M. 2019. How good are pathogenicity predictors in detecting benign variants? *PLoS Comput Biol,* 15**,** e1006481.

NIU, B., YE, K., ZHANG, Q., LU, C., XIE, M., MCLELLAN, M. D., WENDL, M. C. & DING, L. 2014. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics,* 30**,** 1015-6.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C. M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V. S., KODALI, V. K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K. M., MURPHY, M. R., O'NEILL, K., PUJAR, S., RANGWALA, S. H., RAUSCH, D., RIDDICK, L. D., SCHOCH, C., SHKEDA, A., STORZ, S. S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R. E., VATSAN, A. R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M. J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T. D. & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res,* 44**,** D733-45.

PAPAVRAMIDOU, N., PAPAVRAMIDIS, T. & DEMETRIOU, T. 2010. Ancient Greek and Greco-Roman methods in modern surgical treatment of cancer. *Ann Surg Oncol,* 17**,** 665-7.

PELTOMAKI, P., AALTONEN, L. A., SISTONEN, P., PYLKKANEN, L., MECKLIN, J. P., JARVINEN, H., GREEN, J. S., JASS, J. R., WEBER, J. L., LEACH, F. S. & ET AL. 1993. Genetic mapping of a locus predisposing to human colorectal cancer. *Science,* 260**,** 810-2.

PELTOMAKI, P., GAO, X. & MECKLIN, J. P. 2001. Genotype and phenotype in hereditary nonpolyposis colon cancer: a study of families with different vs. shared predisposing mutations. *Fam Cancer,* 1**,** 9-15.

PETERS, U., BIEN, S. & ZUBAIR, N. 2015. Genetic architecture of colorectal cancer. *Gut,* 64**,** 1623-36.

PFEIFER, G. P., YOU, Y. H. & BESARATINIA, A. 2005. Mutations induced by ultraviolet light. *Mutat Res,* 571**,** 19-31.

PICELLI, S., VANDROVCOVA, J., JONES, S., DJUREINOVIC, T., SKOGLUND, J., ZHOU, X. L., VELCULESCU, V. E., VOGELSTEIN, B. & LINDBLOM, A. 2008. Genome-wide linkage scan for colorectal cancer susceptibility genes supports linkage to chromosome 3q. *BMC Cancer,* 8**,** 87.

PLAGNOL, V., CURTIS, J., EPSTEIN, M., MOK, K. Y., STEBBINGS, E., GRIGORIADOU, S., WOOD, N. W., HAMBLETON, S., BURNS, S. O., THRASHER, A. J., KUMARARATNE, D., DOFFINGER, R. & NEJENTSEV, S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics,* 28**,** 2747-54.

POPE, B. J., NGUYEN-DUMONT, T., ODEFREY, F., HAMMET, F., BELL, R., TAO, K., TAVTIGIAN, S. V., GOLDGAR, D. E., LONIE, A., SOUTHEY, M. C. & PARK, D. J. 2013. FAVR (Filtering and Annotation of Variants that are Rare): methods to facilitate the analysis of rare germline genetic variants from massively parallel sequencing datasets. *BMC Bioinformatics,* 14**,** 65.

POWELL, S. M., PETERSEN, G. M., KRUSH, A. J., BOOKER, S., JEN, J., GIARDIELLO, F. M., HAMILTON, S. R., VOGELSTEIN, B. & KINZLER, K. W. 1993. Molecular diagnosis of familial adenomatous polyposis. *N Engl J Med,* 329**,** 1982-7.

RANEY, B. J., CLINE, M. S., ROSENBLOOM, K. R., DRESZER, T. R., LEARNED, K., BARBER, G. P., MEYER, L. R., SLOAN, C. A., MALLADI, V. S., ROSKIN, K. M., SUH, B. B., HINRICHS, A. S., CLAWSON, H., ZWEIG, A. S., KIRKUP, V., FUJITA, P. A., RHEAD, B., SMITH, K. E., POHL, A., KUHN, R. M., KAROLCHIK, D., HAUSSLER, D. & KENT, W. J. 2011. ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res,* 39**,** D871-5.

REED, T. E. & NEEL, J. V. 1955. A genetic study of multiple polyposis of the colon with an appendix deriving a method of estimating relative fitness. *Am J Hum Genet,* 7**,** 236-63.

RIOS GARCIA, M., STEINBAUER, B., SRIVASTAVA, K., SINGHAL, M., MATTIJSSEN, F., MAIDA, A., CHRISTIAN, S., HESS-STUMPP, H., AUGUSTIN, H. G., MULLER-DECKER, K., NAWROTH, P. P., HERZIG, S. & BERRIEL DIAZ, M. 2017. Acetyl-CoA Carboxylase 1-Dependent Protein Acetylation Controls Breast Cancer Metastasis and Recurrence. *Cell Metab,* 26**,** 842-855 e5.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

ROBINSON, J. T., THORVALDSDOTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

ROSENBLOOM, K. R., DRESZER, T. R., LONG, J. C., MALLADI, V. S., SLOAN, C. A., RANEY, B. J., CLINE, M. S., KAROLCHIK, D., BARBER, G. P., CLAWSON, H., DIEKHANS, M., FUJITA, P. A., GOLDMAN, M., GRAVELL, R. C., HARTE, R. A., HINRICHS, A. S., KIRKUP, V. M., KUHN, R. M., LEARNED, K., MADDREN, M., MEYER, L. R., POHL, A., RHEAD, B., WONG, M. C., ZWEIG, A. S., HAUSSLER, D. & KENT, W. J. 2012. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res,* 40**,** D912-7.

ROZEN, S. & SKALETSKY, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* 132**,** 365-86.

RUBIN, A. F. & GREEN, P. 2009. Mutation patterns in cancer genomes. *Proc Natl Acad Sci U S A,* 106**,** 21766-70.

SAMPSON, J. R., DOLWANI, S., JONES, S., ECCLES, D., ELLIS, A., EVANS, D. G., FRAYLING, I., JORDAN, S., MAHER, E. R., MAK, T., MAYNARD, J., PIGATTO, F., SHAW, J. & CHEADLE, J. P. 2003. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet,* 362**,** 39-41.

SCHWARZ, J. M., RODELSPERGER, C., SCHUELKE, M. & SEELOW, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods,* 7**,** 575-6.

SEGUI, N., MINA, L. B., LAZARO, C., SANZ-PAMPLONA, R., PONS, T., NAVARRO, M., BELLIDO, F., LOPEZ-DORIGA, A., VALDES-MAS, R., PINEDA, M., GUINO, E., VIDAL, A., SOTO, J. L., CALDES, T., DURAN, M., URIOSTE, M., RUEDA, D., BRUNET, J., BALBIN, M., BLAY, P., IGLESIAS, S., GARRE, P., LASTRA, E., SANCHEZ-HERAS, A. B., VALENCIA, A., MORENO, V., PUJANA, M. A., VILLANUEVA, A., BLANCO, I., CAPELLA, G., SURRALLES, J., PUENTE, X. S.

& VALLE, L. 2015. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology,* 149**,** 563-6.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research,* 29**,** 308-311.

SHLIEN, A., CAMPBELL, B. B., DE BORJA, R., ALEXANDROV, L. B., MERICO, D., WEDGE, D., VAN LOO, P., TARPEY, P. S., COUPLAND, P., BEHJATI, S., POLLETT, A., LIPMAN, T., HEIDARI, A., DESHMUKH, S., AVITZUR, N., MEIER, B., GERSTUNG, M., HONG, Y., MERINO, D. M., RAMAKRISHNA, M., REMKE, M., ARNOLD, R., PANIGRAHI, G. B., THAKKAR, N. P., HODEL, K. P., HENNINGER, E. E., GOKSENIN, A. Y., BAKRY, D., CHARAMES, G. S., DRUKER, H., LERNER-ELLIS, J., MISTRY, M., DVIR, R., GRANT, R., ELHASID, R., FARAH, R., TAYLOR, G. P., NATHAN, P. C., ALEXANDER, S., BEN-SHACHAR, S., LING, S. C., GALLINGER, S., CONSTANTINI, S., DIRKS, P., HUANG, A., SCHERER, S. W., GRUNDY, R. G., DURNO, C., ARONSON, M., GARTNER, A., MEYN, M. S., TAYLOR, M. D., PURSELL, Z. F., PEARSON, C. E., MALKIN, D., FUTREAL, P. A., STRATTON, M. R., BOUFFET, E., HAWKINS, C., CAMPBELL, P. J., TABORI, U. & BIALLELIC MISMATCH REPAIR DEFICIENCY, C. 2015. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet,* 47**,** 257-62.

SIEBER, O. M., LIPTON, L., CRABTREE, M., HEINIMANN, K., FIDALGO, P., PHILLIPS, R. K., BISGAARD, M. L., ORNTOFT, T. F., AALTONEN, L. A., HODGSON, S. V., THOMAS, H. J. & TOMLINSON, I. P. 2003. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med,* 348**,** 791-9.

SILLARS-HARDEBOL, A. H., CARVALHO, B., BELIEN, J. A., DE WIT, M., DELIS-VAN DIEMEN, P. M., TIJSSEN, M., VAN DE WIEL, M. A., PONTEN, F., MEIJER, G. A. & FIJNEMAN, R. J. 2012. CSE1L, DIDO1 and RBM39 in colorectal adenoma to carcinoma progression. *Cell Oncol (Dordr),* 35**,** 293-300.

SIM, N. L., KUMAR, P., HU, J., HENIKOFF, S., SCHNEIDER, G. & NG, P. C. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res,* 40**,** W452-7.

SINGH, A., SWEENEY, M. F., YU, M., BURGER, A., GRENINGER, P., BENES, C., HABER, D. A. & SETTLEMAN, J. 2012. TAK1 inhibition promotes apoptosis in KRAS-dependent colon cancers. *Cell,* 148**,** 639-50.

SINGH, R., YADAV, V., KUMAR, S. & SAINI, N. 2015. MicroRNA-195 inhibits proliferation, invasion and metastasis in breast cancer cells by targeting FASN, HMGCR, ACACA and CYP27B1. *Sci Rep,* 5**,** 17454.

SLATTERY, M. L. 2004. Physical activity and colorectal cancer. *Sports Med,* 34**,** 239-52.

SPEIR, M. L., ZWEIG, A. S., ROSENBLOOM, K. R., RANEY, B. J., PATEN, B., NEJAD, P., LEE, B. T., LEARNED, K., KAROLCHIK, D., HINRICHS, A. S., HEITNER, S., HARTE, R. A., HAEUSSLER, M., GURUVADOO, L., FUJITA, P. A., EISENHART, C., DIEKHANS, M., CLAWSON, H., CASPER, J., BARBER, G. P., HAUSSLER, D., KUHN, R. M. & KENT, W. J. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res,* 44**,** D717-25.

SPUDICH, G. M. & FERNANDEZ-SUAREZ, X. M. 2010. Touring Ensembl: a practical guide to genome browsing. *BMC Genomics,* 11**,** 295.

STENSON, P. D., MORT, M., BALL, E. V., SHAW, K., PHILLIPS, A. & COOPER, D. N. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet,* 133**,** 1-9.

STREUBEL, B., LAMPRECHT, A., DIERLAMM, J., CERRONI, L., STOLTE, M., OTT, G., RADERER, M. & CHOTT, A. 2003. T(14;18)(q32;q21) involving IGH and MALT1 is a frequent chromosomal aberration in MALT lymphoma. *Blood,* 101**,** 2335-9.

SYNGAL, S., BANDIPALLIAM, P. & BOLAND, C. R. 2005. Surveillance of patients at high risk for colorectal cancer. *Med Clin North Am,* 89**,** 61-84, vii-viii.

THOMPSON, B. A., SPURDLE, A. B., PLAZZER, J. P., GREENBLATT, M. S., AKAGI, K., AL-MULLA, F., BAPAT, B., BERNSTEIN, I., CAPELLA, G., DEN DUNNEN, J. T., DU SART, D., FABRE, A., FARRELL, M. P., FARRINGTON, S. M., FRAYLING, I. M., FREBOURG, T., GOLDGAR, D. E., HEINEN, C. D., HOLINSKI-FEDER, E., KOHONEN-CORISH, M., ROBINSON, K. L., LEUNG, S. Y., MARTINS, A., MOLLER, P., MORAK, M., NYSTROM, M., PELTOMAKI, P., PINEDA, M., QI, M., RAMESAR, R., RASMUSSEN, L. J., ROYER-POKORA, B., SCOTT, R. J., SIJMONS, R., TAVTIGIAN, S. V., TOPS, C. M., WEBER, T., WIJNEN, J., WOODS, M. O., MACRAE, F. & GENUARDI, M. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet,* 46**,** 107-115.

THUTKAWKORAPIN, J., LINDBLOM, A. & THAM, E. 2019. Exome sequencing in 51 early onset non-familial CRC cases. *Mol Genet Genomic Med***,** e605.

THUTKAWKORAPIN, J., PICELLI, S., KONTHAM, V., LIU, T., NILSSON, D. & LINDBLOM, A. 2016. Exome sequencing in one family with gastric- and rectal cancer. *BMC Genet,* 17**,** 41.

TRYKA, K. A., HAO, L., STURCKE, A., JIN, Y., WANG, Z. Y., ZIYABARI, L., LEE, M., POPOVA, N., SHAROPOVA, N., KIMURA, M. & FEOLO, M. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res,* 42**,** D975-9.

UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res,* 40**,** e115.

VALLE, L., HERNANDEZ-ILLAN, E., BELLIDO, F., AIZA, G., CASTILLEJO, A., CASTILLEJO, M. I., NAVARRO, M., SEGUI, N., VARGAS, G., GUARINOS, C., JUAREZ, M., SANJUAN, X., IGLESIAS, S., ALENDA, C., EGOAVIL, C., SEGURA, A., JUAN, M. J., RODRIGUEZ-SOLER, M., BRUNET, J., GONZALEZ, S., JOVER, R., LAZARO, C., CAPELLA, G., PINEDA, M., SOTO, J. L. & BLANCO, I. 2014. New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Human Molecular Genetics,* 23**,** 3506-3512.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J., BANKS, E., GARIMELLA, K. V., ALTSHULER, D., GABRIEL, S. & DEPRISTO, M. A. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics,* 43**,** 11 10 1-33.

VAN DER VELDE, K. J., KUIPER, J., THOMPSON, B. A., PLAZZER, J. P., VAN VALKENHOEF, G., DE HAAN, M., JONGBLOED, J. D., WIJMENGA, C., DE KONING, T. J., ABBOTT, K. M., SINKE, R., SPURDLE, A. B., MACRAE, F., GENUARDI, M., SIJMONS, R. H., SWERTZ, M. A. & INSI, G. H. T. G. 2015. Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Hum Mutat,* 36**,** 712-9.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-51.

WANG, C., FENG, Z., JIANG, K. & ZUO, X. 2017. Upregulation of MicroRNA-935 Promotes the Malignant Behaviors of Pancreatic Carcinoma PANC-1 Cells via Targeting Inositol Polyphosphate 4-Phosphatase Type I Gene (INPP4A). *Oncol Res,* 25**,** 559-569.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

WATERS, L. S., MINESINGER, B. K., WILTROUT, M. E., D'SOUZA, S., WOODRUFF, R. V. & WALKER, G. C. 2009. Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. *Microbiology and Molecular Biology Reviews : MMBR,* 73**,** 134-154.

WEREN, R. D., LIGTENBERG, M. J., KETS, C. M., DE VOER, R. M., VERWIEL, E. T., SPRUIJT, L., VAN ZELST-STAMS, W. A., JONGMANS, M. C., GILISSEN, C., HEHIR-KWA, J. Y., HOISCHEN, A., SHENDURE, J., BOYLE, E. A., KAMPING, E. J., NAGTEGAAL, I. D., TOPS, B. B., NAGENGAST, F. M., GEURTS VAN KESSEL, A., VAN KRIEKEN, J. H., KUIPER, R. P. & HOOGERBRUGGE, N. 2015. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet,* 47**,** 668-71.

WONG, S. H., SUNG, J. J., CHAN, F. K., TO, K. F., NG, S. S., WANG, X. J., YU, J. & WU, W. K. 2013. Genome-wide association and sequencing studies on colorectal cancer. *Semin Cancer Biol,* 23**,** 502-11.

XU, Q. & DUNBRACK, R. L., JR. 2012. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics,* 28**,** 2763-72.

ZHANG, J. X., FU, L., DE VOER, R. M., HAHN, M. M., JIN, P., LV, C. X., VERWIEL, E. T., LIGTENBERG, M. J., HOOGERBRUGGE, N., KUIPER, R. P., SHENG, J. Q. & GEURTS VAN KESSEL, A. 2015. Candidate colorectal cancer predisposing gene variants in Chinese early-onset and familial cases. *World J Gastroenterol,* 21**,** 4136-49.

ZHANG, K., CIVAN, J., MUKHERJEE, S., PATEL, F. & YANG, H. 2014. Genetic variations in colorectal cancer risk and clinical outcome. *World J Gastroenterol,* 20**,** 4167-77.

Karolinska
Institutet